Investigating the transcriptional regulation of *BCL11A* in Triple-negative Breast Cancer

Nicola Tsang Hoi Yee

Department of Pharmacology

University of Cambridge

This thesis submitted for the degree of

Doctor of Philosophy

Lucy Cavendish College

April 2019

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Declaration and specified in the text.

It is not substantially the same as any that I have submitted for any qualification at the University of Cambridge or any other University or similar institution except as declared in the Declaration and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any qualifications at the University of Cambridge or any other University or similar institution except as declared in the Declaration and specified in the text.

It does not exceed the prescribed word limit of 60,000 words exclusive of tables, references and appendices.

List of Collaborations

Mass Spectrometry analyses for the RIME experiment were performed at the proteomics facility at CRUK, CI.

Nicola Tsang Hoi Yee

April 2019

ABSTRACT

Title: Investigating the transcriptional regulation of *BCL11A* in Triple-negative Breast Cancer

Name: Nicola Tsang Hoi Yee

Breast cancer is the most common type of cancer in women worldwide. It can be classified into several subtypes based on the histological expression of Estrogen, Progesterone and HER-2 receptors. The most aggressive subtype of breast cancer is called Triple-negative breast cancer (TNBC) and is clinically defined by the absence of the three receptors mentioned above. These TNBC patients have the poorest prognostic outcome among the subtypes and highest mortality rate due to the lack of targeted treatments.

A recent study has shown the expression of the transcription regulator *BCL11A*, to be upregulated specifically in TNBC compared to the other subtypes of breast cancer. In addition, *BCL11A* has also been experimentally demonstrated to be an oncogene in these aggressive tumours, thus making it an interesting target for new drug developments. However, little is known about how *BCL11A* is regulated at the transcriptional level in breast cancer. To tackle this problem, this project implements a novel proteomic approach utilising the Clustered regularly interspaced short palindromic repeats – CRISPR associated 9 (CRISPR-Cas9) technology, to identify putative transcriptional regulators of *BCL11A* in TNBC.

The first step of this project involves the identification of regulatory sites on *BCL11A* that display high transcriptional regulations. This was achieved using the CRISPR-Cas9 knock-out approach. We then employed the mutated catalytically inactive Cas9 (dCas9) to target these sites and to pull down proteins in close proximity. This novel approach led to the identification of DEK, SSRP1 and PSIP1 as potential regulators of *BCL11A*'s expression. Experimentally, I found knocking-down *DEK*, *SSRP1* and *PSIP1* resulted in a reduction in *BCL11A*'s expression in TNBC. These results highlight a novel approach for the identification of

transcription regulators of *BCL11A*'s in TNBC and potential targets for therapeutic interventions for these patients.

This dissertation is dedicated to my most loving mother and grandfather.

ACKNOWLEDGEMENTS

I have been very fortunate to have spent the past four years in the Department of Pharmacology, University of Cambridge, under Dr. Walid Khaled. Firstly, I would like to thank Walid for giving me the opportunity to be his student, and for giving me his guidance, encouragement, and support throughout this project. I am also very grateful for his supervision and great passion for science that carried me through this intense but rewarding journey.

I would also like to thank all of the past and present WTK laboratory members for all the laughter during and outside of work. I have enjoyed myself very much! In particular, I would like to thank Dr. Sara Pensa and Dr. Kyren Lazarus for introducing and helping me to various techniques, especially FACS sorting and RNA interference; for their help, friendly suggestions, stimulating conversations, proofreading part of this thesis, and providing insightful feedback. I would also like to extend my thanks to Dr. Maria Francesca Santolla and Dr. Mark Waterhouse for their advice and help with the Co-immunoprecipitation experiments. And of course, I would like to thank the other students, Elisabetta Zambon, Fazal Hadi, Rosie Ugur, Karsten Bach, Jessica Scheel, and Marija Zarocsinceva for providing such a friendly workplace and all their support over the past four years. Not to forget, thank you Flavia Peci for all the enjoyable conversations in the office. I wish you all the very best in the future!

I would like to further extend my thanks to my friends in the Department of Pharmacology, especially Peace Uredojo, Winnie Yeung, Dewi Safitri Chaniago, Sam Chakrabarti, and Rosie Water for the coffee breaks and delightful chats.

Next, I would like to give my thanks to all my friends in the UK and around the world, especially Stephanie Lau, Celia Lau, Carrie Hui, Felicity Ng, Jennifer Wong, Karen Chan, Kathryn Wu, Weiling Yuan for their friendship and support; and Charine Lo, Lilian Fong, Christine Wong, and Yining Neo for flying half-way across the globe to visit me. I would also like to thank Saumya Srivastava, Stephen Siu, Vivian Sze-To, Lee WenYao, and Wade Suen for all the long phone calls,

coffee chats, and their encouragement. Apart from them, I would like to thank Adrian Chung, Keith Chan, Wilson Suen, and the Legends for all the good times we had and for helping me pull through these four years.

I would also like to give my special thanks to Auntie Sylvia for her constant encouragement and for keeping me calm; and Ryan Ng for all the late nights, warm encouragement, proofreading parts of this thesis, always putting a smile on my face, and putting up with me.

Finally, I would like to give me biggest thanks to my family for their unconditional love and support, being by my side all the time, and standing through thick and thin. I will never forget what we have gone through in the past four years; every single thing made me stronger than ever before and shaped me into who I am today.

"I was taught that the way of progress was neither swift nor easy." - Marie Curie

CONTENTS

DECLARATION ..... **II**

ABSTRACT ..... **III**

ACKNOWLEDGEMENTS ..... **VI**

LIST OF FIGURES ..... **XV**

LIST OF TABLES ..... **XIX**

ABBREVIATIONS ..... **XXII**

CHAPTER 1 Introduction **2**

1.1 Introduction 3

1.2 Breast cancer 3

1.2.1 Anatomy of the breast and the mammary gland 4

1.2.2 Classifying breast cancer subtypes 5

1.2.2.1 Clinical decisions 7

1.2.2.2 Subtyping with molecular markers 7

1.2.2.3 Basal-like and Triple-negative subtype 8

1.3 BCL11A 11

1.3.1 Expressions and roles in biological systems 12

1.3.1.1 Roles in the immune system 12

1.3.1.2 Roles in the neural system 14

1.3.1.3 Roles in Lung cancer 14

1.3.1.4 Roles in Triple-negative breast cancer 17

1.4 Transcriptional regulation 18

1.4.1 Transcription factors 18

1.5 Proteomics studies 20
1.5.1 Traditional methods to study proteomics and Rapid Immunoprecipitation of endogenous proteins with mass spectrometry (RIME) 21
1.5.2 Applications of RIME 22
1.6 Genetic engineering and genome editing 24
1.6.1 The CRISPR-Cas9 system 24
1.6.1.1 Mechanism of the system 24
1.6.1.2 Comparison between traditional genome editing techniques and CRISPR-Cas9 25
1.6.1.3 Designing components for the CRISPR-Cas9 system 25
1.6.1.4 Applications of the CRISPR-Cas9 system 28
1.7 RNA interference technology 29
1.8 Aims and outlines of this project 31
CHAPTER 2 Materials and Methods 34
2.1 Methods 35
2.1.1 Cloning 35
2.1.1.1 Guide RNA design and cloning into PiggyBac vector 35
2.1.1.2 Knocking-out *BCL11A* with CRISPR-Cas9 vectors 37
2.1.1.3 Designing shRNA primers and cloning into PiggyBac vector . 37
2.1.2 Tissue culture techniques 40
2.1.2.1 Culturing cell lines 40
2.1.2.2 Cell transfections 40
2.1.2.3 Proliferation assay 43
2.1.2.4 Colony forming assay 43

2.1.3 Reverse transcriptase (RT)- polymerase chain reaction (PCR) 43
2.1.4 Protein extraction 44
2.1.5 Western blot 44
2.1.5.1 Optimising Western blot conditions for antibodies – Protein DEK (DEK), Facilitates chromatin transactions (FACT) complex subunit SSRP1 (SSRP1) and PC4 and SFRS1 interacting protein 1 (PSIP1) 47
2.1.6 RIME 47
2.1.6.1 Optimising sonication and determining reverse crosslinking conditions 50
2.1.6.1.1 MDA-MB-231 cells 50
2.1.6.1.2 HCC1569 cells 51
2.1.7 Chromatin Immunoprecipitation 51
CHAPTER 3 Designing a novel system to target *BCL11A*'s genomic regions with the CRISPR-Cas9 system 55
3.1 Introduction 56
3.2 Results 58
3.2.1 The determination of which *BCL11A* genomic regions to study and the design of guide RNAs for expression plasmids 58
3.2.2 The importance of *BCL11A*'s genomic targets for cell survival 58
3.2.3 The generation of stable cell lines to express gRNAs and dCas9 60
3.2.4 The validation of the induction and expression of dCas9 using the Tet-On system 65
3.2.4.1 Determination that dCas9 is inducible and its impact on *BCL11A* expression investigated 65

3.2.5 Determination that dCas9 binds specifically at its corresponding targets 66
3.3 Discussion 70
3.3.1 The evaluation of the importance of chosen target sites 70
3.3.2 The validation of components of the experimental design..... 70
CHAPTER 4 Proteomic analysis..... 72
4.1 Introduction 73
4.2 Results 75
4.2.1 Formation of a reverse crosslink for RIME to visualise sizes of sonicated chromatin 75
4.2.2 The analysis of protein pull-down from the RIME dataset 75
4.2.3 The validation of protein pull-down from RIME 87
4.2.3.1 The expression of candidates in TNBC cell lines 87
4.2.3.2 Confirmation by ChIP that the proteins bind at the target sites 87
4.2.3.3. The evaluation of any bias in target binding by DEK, SSRP1 and PSIP1..... 95
4.3 Discussion 97
4.3.1 Analysis of RIME pull-down..... 97
4.3.2 Evaluation of the binding of candidates in TNBC cell lines 100
4.3.3 Do DEK, SSRP1 and PSIP1 form a complex? 101
CHAPTER 5 Investigating the functional effects of DEK, SSRP1 and PSIP1 using shRNA knock-down 103
5.1 Introduction 104
5.2 Results 105
5.2.1 The design of primers for the shRNA-expression plasmid 105

5.2.2 The generation of stable cell lines that express shRNA 105
5.2.3 The validation of shRNA knock-down 105
5.2.4 Exploration of the effects of shRNA knock-down..... 112
5.2.4.1 The effects of shRNA knock-down on expression of *BCL11A* 112
5.2.4.2 The functional effects of shRNA knock-down 112
5.2.5 Candidates work together to regulate expression of *BCL11A* 115
5.2.5.1 Double shRNA knock-down affects expression of *BCL11A* .. 118
5.3 Discussion 126
5.3.1 Evaluation of the experimental design 126
5.3.2 Functional effects of shRNAs..... 126
5.3.3 The effect of double shRNA knock-down..... 128
5.3.4 Proposed mechanism for regulation of expression of *BCL11A* by DEK, SSRP1 and PSIP1 129
5.4 Conclusion..... 130
CHAPTER 6 Discussion, future directions, and conclusion 132
6.1 Discussion 133
6.1.1 Rationale behind the research aim and Achievements..... 133
6.1.2 Evaluating the CRISPR-Cas9 system 134
6.1.3 Evaluating the experimental design and proteomic analysis from the RIME experiment 136
6.1.4 Implications of RIME pull-down..... 138
6.2 Future directions 140
6.3 Conclusion..... 143

REFERENCES 144

APPENDIX A 163

LIST OF FIGURES

CHAPTER 1

Figure 1.1: The cycle of mammary gland at puberty, maturity, during and at the end of pregnancy……...……...…….6

Figure 1.2: Structure of the *BCL11A* gene and the four isoforms……..…13

Figure 1.3: Structure of the β- globin gene clusters…….……..…..…....16

Figure 1.4: An overview of the processes in transcription……..……19

Figure 1.5: The major steps in RIME…….....……...…...…23

Figure 1.6: The CRISPR-Cas9 system……...…….…….....26

Figure 1.7: The two pathways to repair double-strand breaks……..…...27

Figure 1.8: The mechanism adapted by catalytically inactive Cas9 (dCas9)…..30

Figure 1.9: The mechanism of the RNA interference technique……....…32

CHAPTER 2

Figure 2.1: Optimisations for the conditions of western blot for primary antibodies, DEK, SSRP1 and PSIP1……...….....49

Figure 2.2: Optimising sonication and determining reverse crosslinking conditions for MDA-MB-231 and HCC1569 cells…….……...…….52

CHAPTER 3

Figure 3.1: Designing gRNA complementary to designated sites…….....59

Figure 3.2: Deleting Sites A and B with Cas9 expression plasmids……...61

Figure 3.3: gRNA- and dCas9-expressing vectors and gene delivery…….....63

Figure 3.4: Optimisations of co-transfections ……..……...……..64

Figure 3.5: Effect of *BCL11A* expression with dCas9 induction for 24 and 48 hours……..…….67

Figure 3.6: ChIP on samples with gRNA for Site A showed specific dCas9 binding…….....……68

Figure 3.7: ChIP on samples with gRNA for Site B showed specific dCas9 binding…….……...……...69

CHAPTER 4

Figure 4.1: Experimental work flow for RIME……..…....74

Figure 4.2: Interpreting the RIME dataset……..……..…….…...76

Figure 4.3: cBioportal correlating mRNA expression with PAM_{50} and *BCL11A* expression……...…….……...…….....88

Figure 4.4: Protein expression of candidates on a panel of breast cancer cell lines……...……..89

Figure 4.5: ChIP experiment results, showing candidates binding in MDA-MB-231 cells……...……...…….....91

Figure 4.6: Specificities of candidates' bindings in MDA-MB-231 cells.....92

Figure 4.7: ChIP experiment results, showing candidates bindings in HCC1569 cells……...…..……93

Figure 4.8: Specificities of candidate bindings in HCC1569 cells……..….....94

CHAPTER 5

Figure 5.1: The design of shRNA primers and the schematic of the experimental plan……...…….106

Figure 5.2: Validation of the expression of *DEK* gene knock-down with RT-PCR…….……108

Figure 5.3: Validation of the expression of *SSRP1* gene knock-down with RT-PCR……109

Figure 5.4: Validation of the expression of *PSIP1* gene knock-down with RT-PCR……110

Figure 5.5: Protein expression of DEK, SSRP1 and PSIP1 in shRNA samples……111

Figure 5.6: Evaluation of the effects of shRNA knock-down on expression of *BCL11A*……113

Figure 5.7: Evaluation of the effects of shRNA knock-down on global gene expressions……114

Figure 5.8: Functional studies of the knock-down of DEK, SSRP1 and PSIP1……116

Figure 5.9: The schematic of the experimental plan for double shRNA knock-down……117

Figure 5.10: Validation through RT-PCR of the knock-down of expression of *DEK* in double knock-down samples……119

Figure 5.11: Validation through RT-PCR of the knock-down of expression of *SSRP1* in double knock-down samples……120

Figure 5.12: Validation through RT-PCR of the knock-down of expression of *PSIP1* in double knock-down samples……121

Figure 5.13: Protein expression for DEK, SSRP1 and PSIP1 in double knock-down samples……122

Figure 5.14: Evaluation of the effects of double shRNA knock-down on expression of *BCL11A* and *PPP2R2D*……124

Figure 5.15: Protein expression for BCL11A in double knock-down samples……125

Figure 5.16: Proposed mechanisms of regulation of *BCL11A*'s expression……..……..…131

LIST OF TABLES

CHAPTER 1

Table 1.1: Intrinsic subtyping by Perou et al and the characteristics of each subtype……..……..……..……9

Table 1.2: Examples of *BCL11A* binding partners and their roles in relation to the functions of *BCL11A*……...…..……...……..15

CHAPTER 2

Table 2.1: The gRNA sequences designed for targeting *BCL11A* with dCas9 …….……..... 36

Table 2.2: The gRNA sequences designed for knocking out Sites A and B ……...…..38

Table 2.3: The shRNA sequences obtained for candidates knock-down ……...……...…….……. 39

Table 2.4: Summary of the culture conditions for the breast cancer cell lines used……...……...……41

Table 2.5: Summary of cell transfections performed……...…...42

Table 2.6: List of the expression of genes quantified in RT-PCR……..…45

Table 2.7: The components for making Acrylamide gels in western blot experiments ……...…….,…….……...… 46

Table 2.8: The antibodies used in western blot experiments……...…..48

Table 2.9: The primer sequences used in ChIP-qPCR experiments……..…....54

CHAPTER 4

Table 4.1: Summary of top 20 proteins from the RIME analysis after eliminating common proteins between IgG and triplicates between two target sites ……...……... 77

Table 4.2: Summary of top 20 proteins from the RIME analysis after eliminating common proteins between IgG and triplicates at Site A……..…78

Table 4.3: Summary of top 20 proteins from the RIME analysis after eliminating common proteins between IgG and triplicates at Site B……..……79

Table 4.4: Summary of top 20 proteins from the RIME analysis after the of filtering for proteins with nuclear expressions between the two targets ……...…….……... 81

Table 4.5: Summary of top 20 proteins from the RIME analysis after filtering for proteins with nuclear expressions at Site A……..,……82

Table 4.6: Summary top 20 proteins from the RIME analysis after filtering for proteins with nuclear expressions at Site B……..…..…83

Table 4.7: Summary of top 20 proteins from the RIME analysis after referring to the Crapome database between Sites A and B……..……...…….84

Table 4.8: Summary of top 20 proteins from the RIME analysis after referring to the Crapome database at Site A……..…....……85

Table 4.9: Summary of top 20 proteins from the RIME analysis after referring to the Crapome database at Site B……...……86

Table 4.10: Summary of statistical analysis of binding bias by DEK, SSRP1 and PSIP1 in MDA-MB-231 and HCC1569 cell lines……...……96

Table 4.11: Summary of known cellular functions and roles of DEK, SSRP1 and PSIP1 in breast cancer……...…..……98

APPENDIX A

Appendix table A1: List of the proteins after the elimination of common proteins between IgG and the triplicates at Site A……..…...164

Appendix table A2: List of the proteins after the elimination of common proteins between IgG and triplicates of Site B…….166

Appendix table A3: List of the proteins remained after the elimination of common proteins between the IgG control and the triplicates of Sites A and B ……...……167

Appendix table A4: Proteins pulled-down from Site A with nuclear expressions……...…….168

Appendix table A5: Proteins with nuclear expressions pulled-down in RIME at Site B…….……169

Appendix table A6: Common proteins between Sites A and B with nuclear expressions……..……...170

Appendix table A7: Proteins with less than 100 studies from the Crapome database at Site A…….…….171

Appendix table A8: Proteins after the analysis with Crapome study at Site B….....172

Appendix table A9: Common proteins from the Crapome database with less than 100 studies between Sites A and B ……....…….173

ABBREVIATIONS

2D-PAGE	two-dimensional gel electrophoresis
ADP	Adenosine diphsphate
AMBIC	Ammonium hydrogen carbonate
AML	Acute myeloid leukaemia
APEX	Apurinic/ Apyrimidinic Endodeoxyribonuclease
AR	Androgen receptor
ASD	Autism spectrum disorder
β	Beta
BLBC	Basal-like breast cancer
BCL11A	B-cell lymphoma/ leukaemia 11A
BCL11B	B-cell leukaemia 11B
BSA	Bovine serum albumin
C2H2	Cys2His2
cDNA	Complementary DNA
CHD8	Chromodomain helicase DNA binding protein 8
ChIP	Chromatin Immunoprecipitation
CK2	Caesin kinase 2
Co-IP	Co-immunoprecipitation
Crapome	Contaminant repository of affinity purification
CRISPR	Clustered regularly interspaced short palindromic repeats
CRISPR-Cas9	CRISPR- CRISPR associated protein 9
CRISPRi	CRISPR interference
CSH	Control shRNA
CTU1/2	Cytosolic thiouridylase protein 1/2
δ	Delta
DEK	Protein DEK
DMBA	Dimenthylbenz[a]anthracene
DMEM	Dulbecco's modified Easgle's medium
DMEM/F12	DMEM: Nutrient Mixture F12
DNA	Deoxyribonucleic acid
dCas9	Catalytically inactivate Cas9

dsRNA	Double stranded RNA
DTT	Dithiothreitol
ε	Embryonic
E2f4	E2F transcription factor 4
E.coli	Escherichia coli
EDTA	Ethylenediaminetetraacetic acid
EGFR	Epidermal growth factor receptor
ER	Estrogen receptor
ERBB2	Erb-B2 receptor Tyrosine kinase 2
ECL	Enhanced chemiluminescence
EGFP	Enhanced green fluorescent protein
ELP3	Elongator complex protein 3
EMT	Epithelial-mesenchymal transition
FACS	Fluorescence-activated cell sorting
FACT	Facilitates chromatin transactions
FISH	Fluorescence *in situ* hybridisation
γ	Gamma
FBS	Fetal bovine serum
FRET	Fluorescence resonance energy transfer
G1/S	Gap 1/ Synthesis
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GREB1	Growth regulating Estrogen receptor binding 1
Ham/F12	Hasm's F12 nutrient mixture
Hb	Haemoglobin
HbA	Hb adult
HbF	Hb Foetal
HCl	Hydrochloric acid
HDR	Homology directed repair
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesullfonic acid
HER	Human epidermal growth factor
HS1	Hamatopoetic lineage cell-specific protein 1
HSC	Haematopoietic stem cell
IHC	Immunohistochemistry
IRES	Internal ribosome entry site

ITR	Inverse terminal repeats
LCR	Locus control region
LEF1	Lymphoid-enhancer binding factor 1
LUSC	Lung squamous cell carinoma
KOH	Potassium salt
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
MEK/ERK	Ras-Raf-MEK-ERK
mRNA	Messenger ribonuclease acid
NaCl	Sodium Chloride
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NHEJ	Non-homologous end joining
NSCLC	Non-small cell lung cancer
p53	Tumour protein p53
p63	Tumour protein p63
PAM_{50}	Prosigna breast cancer prognostic signature assay
PAM	Protospacer-adjacent motif
PARP	Poly Adenosine diphosphate (ADP)-ricose polymerase
PB	PigggBac
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PDX	Patient derived xenograft
PI3K	Phosphoinositide 3- kinase
PLA	Proximity ligation assay
PPP2R2D	Protein phosphatase 2 regulatory subunit B delta
PR	Progesteron receptor
PS	Penicillin Streptomycin
PSG	Penicillin Streptomycin Glutamine
PSIP1	PC4 and SFRS1 interacting protein 1
PTEN	Phosphatase and tensin homolog
PVDF	Polyvinylidene difluoride
qPCR	Quantitative polymerase chain reaction
Rb1	Retinoblastoma protein 1

RIME	Rapid Immunoprecipitation of endogeneous proteins with mass spectrometry
RIPA	Radioimmunoprecipitation assay
RISC	RNA-induced silencing complex
RMPI	Rosewell park memorial institute 1640 medium
RNA	Ribonuclease acid
RNAi	RNA interference
RNAP	RNA polymerase
RT	Reverse transcriptase
RT-qPCR	Reverse transcription-qPCR
rTtA	Recombinant tetracycline-controlled transcription factor
Setd4	SET domain-containing protein 4
sgRNA	Single guide RNA
SILAC	Stable isotope labelling by ammino acids
siRNA	Small interfering RNA
shRNA	Short hairpin RNA
SNP	Single nucleotide polymorphism
SSRP1	FACT complex subunit SSRP1
STRING	Search tool for the retrieval of interacting genes/proteins
SUPT16	FACT complex subunit SPT16
TALENs	Transcription-activator-like effector nucleases
TBS	Tris-buffered saline
TBS-T	TBS-Tween 20
TCGA	The Cancer Genome Atlas
TE	Tris-EDTA
TEB	Terminal end buds
TNBC	Triple-negative breast cancer
TraccRNA	Trans-encoded small RNA
TRE	Tet reponse element
tRNA	Transfer RNA
UCSC	University of California Santa Cruz
UTR	Untranslated region
UK	United Kingdom
ZFNs	Zinc finger nucleases

CHAPTER 1

Introduction

1.1 Introduction

Breast cancer is the most common type of cancer in women residing in the United Kingdom (UK) [1]. It has been estimated that 1 in 7 women will have breast cancer some time in their life [2]. In order to understand this disease, researchers investigate breast cancer with the aim to understand the molecular underlying and to develop ways to tackle these causes. These have been made possible thanks to advancement in technology, which allows us to classify breast cancer patients clinically under the umbrella of subtypes characterised by specific gene and protein expressions. Broadly speaking, breast cancer can be classified into two subtypes: one expressing hormonal receptors and another without. These are in turn further classified into intrinsic subtypes through the study of gene expression profiles.

The hormonal receptors negative subtype is called Triple-negative breast cancer (TNBC). It does not express receptors for the hormones Estrogen and Progesterone, nor the Human epidermal growth factor (HER)-2, thus making conventional treatments which target these receptors unsuitable for treating TNBC patients and posing a challenge to find new molecular targets for drug development. Studies of this subtype have led to the identification of B-cell lyphoma/ leukaemia 11A (*BCL11A*) as an oncogene and a potential drug target. However, its regulation in TNBC remained to be determined. Therefore, this project aims to investigate the regulation of *BCL11A* in TNBC using the Clustered regularly interspaced short palindromic repeats (CRISPR) – CRISPR associated protein 9 (CRISPR-Cas9) system and proteomic approach. Chapter 1 will give an overview of present knowledge in the field of breast cancer, *BCL11A*, and the research techniques used in this project.

1.2 Breast cancer

Breast cancer is globally the most common type of cancer in females, contributing to 15% of new cancer cases in 2008 [3]. Breast cancer is also the most common

type of cancer in women since 1997 [1]. In 2018 alone, two million new cases were diagnosed in the UK [4]. One contributing factor to an increase in the number of diagnosis in the UK relative to 1990s is that women between the ages of 50 and 70 now receive screening every three years [5]. It has been estimated that one person is diagnosed every ten minutes per day [6]. As a result of this good national screening programme, the rate of mortality has been reduced and the survival rate has doubled in the past forty years. Despite this, the rise of new cases remained steady. A cancer research organisation has predicted that one in seven women would develop breast cancer during their lifetime [2]. A wide range of risk factors have been identified, including aging, excessive alcohol consumptions, obesity, increasing breast density and drugs such as hormone replacement. In addition, being of certain ethnicities, such as African-descendent American, could also increase susceptibility to breast cancer, as demonstrated in the Carolina Breast cancer study [7]. Cancer starts by inheriting or acquiring genetic aberrations accumulated through life. These mutations can lead to changes in gene expressions and disruptions of molecular components or cellular pathways.

1.2.1 Anatomy of the breast and the mammary gland

The breast consists of various cell types including the adipose cells and the epithelial duct lobular units. These duct lobular units ultimately function by producing milk at the lobules and secreting it through ducts that connect the lobules to the nipples for new-born infants. These duct lobular units are also called the mammary gland.

The mammary gland is unique in that most of its development occurs in adulthood and it only reaches maturity during pregnancy under the influence of hormones [8] [9] [10]. Its development is arrested after birth and restarts again when a female reaches puberty. Figure 1.1 describes the changes in the mammary gland at puberty, during and after pregnancy. During puberty, Estrogen and Progesterone initiate ductal growth into the fat pad and these growing ducts have cap-like structures at the tips called the Terminal end buds (TEB) [11] [12] [13]. These TEB are areas with high proliferation rates that can fill the fat pad with primary and secondary

branches [14]. In addition, side branching also occurs in the matured virgin. The mammary gland development then reaches a halt and continues during pregnancy, when a change in hormonal signalling initiates tertiary branching and the TEB develops into milk-producing alveoli for nourishing infants [15]. The mammary gland is a bilayer epithelium consisting of two cell types: luminal and basal cells. The luminal cells face the inner side of the ducts and alveolar while basal cells, which is also called myoepithelial cells, line and face the outer side [16]. These myoepithelial cells are responsible for ejecting milk from the alveolar into the ducts during lactation. Upon the loss of demand for milk from the new-born, the mammary gland undergoes a series of cell death events and tissue remodelling called involution [17] [18] [19] . This returns the gland to the pre-pregnant stage where primary and secondary branching are still missing.

The different cell types in the mammary gland are thought to be the reason for the heterogeneity in breast cancer, as the various subtypes of breast cancer could originate from different cell types in the mammary gland [20]. I will now give an overview of the traditional and current methods for classifying breast cancer subtypes.

1.2.2 Classifying breast cancer subtypes

Traditionally, pathologists use histological analysis to diagnose breast cancer subtypes with established markers: Estrogen receptor (ER), Progesterone receptor (PR), and Her-2 receptor. Presence of the first two traditional hormonal receptors is determined by Immunohistochemistry (IHC). As for the third, amplification of the Her-2 coding gene, Erb-B2 receptor Tyrosine kinase 2 (*ERBB2*), is determined by Fluorescence *in situ* hybridisation (FISH). The identification of these markers in breast cancer samples has also led to the classification of patients into four main subtypes: ER-positive, PR-positive, Her-2 overexpressed, and Triple-negative breast cancer, the last of which does not express any receptors nor has Her-2 amplified. Apart from the identification of subtypes, histological staining also revealed the grading of aggressiveness based on morphological features of differentiation and proliferation [21].

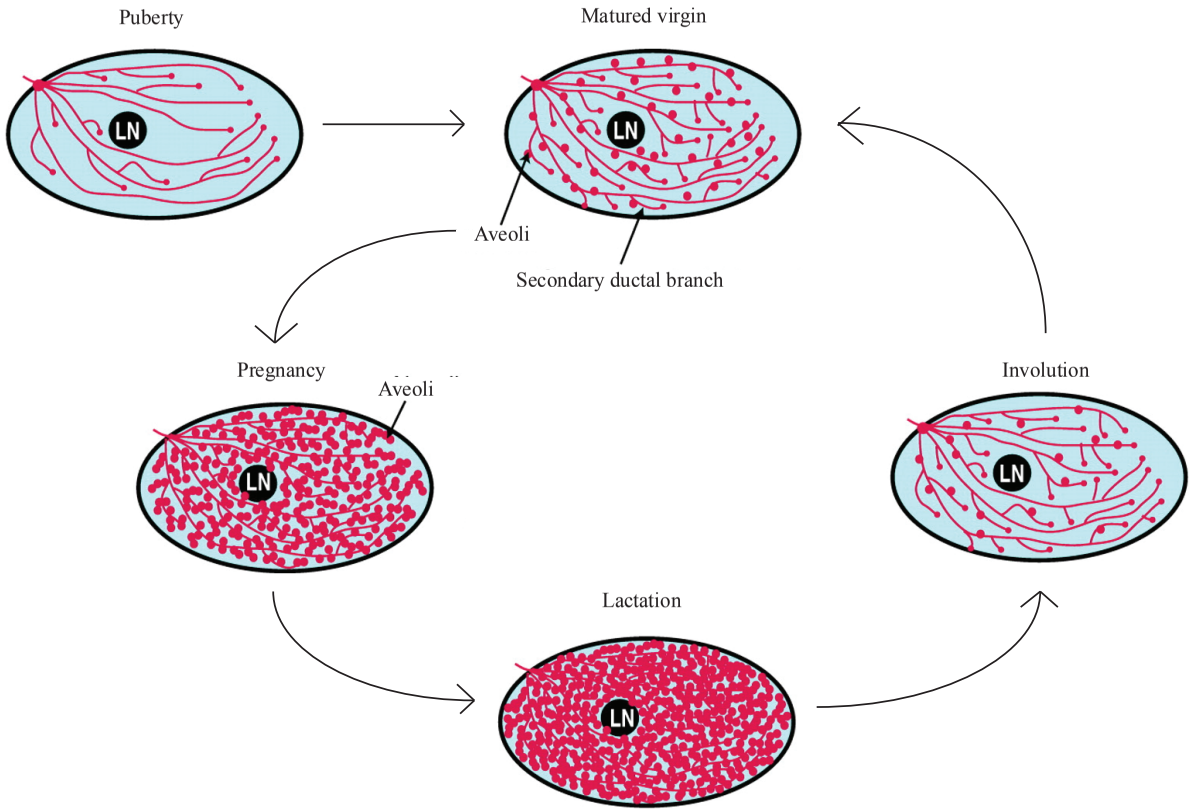


Figure 1.1: The cycle of mammary gland at puberty, maturity, during and at the end of pregnancy.

At puberty, the mammary gland initiates ductal growth that invades into the fat pad with cap-like structures called Terminal end buds (TEB) at the tip. In the matured virgin gland, primary and secondary ductal structures fill that fat pad with side branches, which during pregnancy differentiate into alveolar. These alveolar buds then produce milk during lactation. At the end of lactation, the lack of demand for milk causes the mammary gland to undergo involution and returns the mammary gland to the pre-pregnant stage. (Modified from [22])

1.2.2.1 Clinical decisions

Both subtyping and grading systems from histological staining are thought to provide prognostic information and help to predict clinical outcomes. Thus, medical practitioners have been utilising these results when making decisions on the best treatment available for breast cancer patients. Traditional anti-hormonal treatments were often administered to treat ER- or PR-positives. This drug class works by inhibiting hormones, ER or PR, from binding to their receptors and thus stopping cell growth. Tamoxifen was the first anti-hormonal drug developed among this drug class and is prescribed for patients with ER-positives [23]. Apart from anti-hormonal treatments, Aromatase inhibitors were also prescribed for ER-positives. These work by blocking the enzyme, Aromatase, from converting Androgen into Estrogen [24]. This drug class is prescribed preferentially to post-menopausal patients due to changes in the source of Estrogen; the female sex hormone Estrodiol is the primary form of endogenous Estrogen in women before menopause, while Androgen becomes the predominant source in post-menopausal patients [25]. In contrast, the main treatment for Her-2 positive patients is monoclonal antibodies which bind to the Her-2 receptor and prevent activation of the receptor and its downstream kinase cascades.

None of the drug classes described above is prescribed for TNBC patients due to the lack of hormonal and Her-2 targets. Therefore, chemotherapy is the only option available for these patients and the characterisation with IHC approaches are not accurate enough to meet the subtype's need [26] [27]. Not helping the case is the common phenomenon that regardless of subtype, outcomes of treatments often differ from what would be expected based on observed clinical behaviours. This urges the development of new, well-defined makers for classification and therapeutic approaches.

1.2.2.2 Subtyping with molecular markers

Advancements in diagnostic techniques have led to further subtyping of breast cancer, identification of new drug targets, and better prognostic predictions. In

2000, Perou et al. [28] proposed to classify subgroups by microarray. These gene expression profiles have deepened the understanding of breast cancer heterogeneity. Moreover, it has revolutionised breast cancer subtyping by introducing five 'intrinsic subtypes': Luminal A, Luminal B, Her-2-enriched, Basal-like and normal. An additional subtype, Claudin-low, also became a member in 2007 [29]. Despite their names, overlaps were found between subtypes identified from microarrays and IHC. These are presented in Table 1.1.

1.2.2.2.1 Heterogeneity

Breast cancer heterogeneity is not only observed in the expression of receptors but also in gene expression profiles. The diversity of gene expressions can be interpreted as spatial or temporal control to adapt their microenvironment, and mutation of gene expressions could lead to metastasis. Despite such heterogeneity, the histological detection of ER remains the main marker for determining clinical treatments for patients.

In order to develop more targeted therapeutic strategies, there is a need for a classification of breast cancer that captures heterogeneity better. For that purpose, studies have integrated 'intrinsic subtypes', genomic aberrations, and clinical outcomes. The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) study combines the copy numbers and gene expression profiles from 2000 patients [31]. METABRIC also contributed to The Cancer Genome Atlas (TCGA) dataset which shows heterogeneity at the genetic and molecular levels [32]. Information from the TCGA database is often expressed as the Prosigna breast cancer prognostic signature assay (PAM_{50}) which classifies the intrinsic subtypes by analysing the co-expression of 50 mRNAs [27].

1.2.2.3 Basal-like and Triple-negative subtype

Basal-like breast cancer (BLBC) resembles the morphology of the basal cells of the mammary gland. This subtype is often classified using microarray and IHC techniques with high sensitivity against basal elements [33], such as cytokeratin 5, 6, 14 [34], 17 [28], P-cadherin [35] or Epidermal growth factor receptor (EGFR)

Intrinsic subtypes	Characterisations
Luminal A	ER- /PR- positives, low proliferation rate, best prognosis
Luminal B	ER- &/ PR- positives, higher proliferation rate, less responsive to hormonal treatments, worse prognosis than Luminal A
Her2-enriched	Amplified *ERBB2* or genes from the same chromosome, high grade
Basal-like	Markers of the breast's basal cells, aggressive, poor prognosis
Claudin-low	Similar gene expressions as mammary stem cells
Normal	Markers resemble normal breast tissues

Table 1.1: Intrinsic subtyping by Perou et al and the characteristics of each subtype.

Table 1.1 presents the overlaps of subtypes identified using microarrays, IHC methods, and intrinsic subtypes. In addition, it also shows certain characteristics of subtypes, such as differences in the proliferation rate, prognosis, and resemblance to the breast tissues. [26] [29] [30]

[33] [36]. BLBC contributes to 20% of diagnosed breast tumour cases [3] and shares a lot of features with TNBC. 80% of TNBC cases were identified as BLBC, since both lack hormonal receptors for ER and PR and the amplification of Her-2 from histology analysis. They are also generally high graded tumours with elevated proliferation rates from dysregulating pathways or components of the cell cycle [37]. For example, gene expression analysis detected high levels of Cyclin E1 which promotes Gap 1/ Synthesis (G1/S) phase progression in the cell cycle [38] [39]; loss of Phosphatase and tensin homolog (PTEN) which regulate Phosphoinositide 3- kinase (PI3K) activities [27]; and reduced expression in tumour suppressor retinoblastoma protein 1 (Rb1), resulting in the upregulation of cyclin-dependent kinase I, p16 [37] [40] [41]. Mutations in tumour suppressor Tumour protein p53 (p53) are also commonly found in BLBC [41].

Apart from the molecular similarities between BLBC and TNBC, the two share clinical features including young age (<50 years old), higher susceptibilities in African-Americans, poor prognosis with recurrences in three years after initial treatment [3], and a higher mortality rate than the rest within a five year period [42]. Their aggressiveness can also be seen from overexpressed genes from the Ras-Raf-MEK-ERK (MEK/ ERK) pathway, which promotes cellular invasion, and high levels of associated genes in the epithelial-mesenchymal transition (EMT) mechanism from Sarrió's study [43]. EMT starts by changing the expression of cell-surface markers and cell-cell adhesion molecules, followed by overexpression of proteins responsible for remodelling the extracellular matrix [43]. These simultaneous events allow tumour cells to detach from their original sites and colonise elsewhere in the body. An eight year clinical study [44] found a higher proportion of TNBC patients to have experienced metastasis, while a median of one year survival in metastased TNBC patients was reported by the Abramson group [45]. Furthermore, TNBC was found to have metastatic potentials to specific tissues, such as the central nervous system [44], liver, and the brain [33].

1.2.2.3.1 Developing treatments for TNBC patients

Despite our knowledge of this subtype, treatments are still mostly limited to chemotherapy and the high mortality rate remains unsolved. One way to overcome these is to repurpose existing drugs to treat TNBC patients. An example would be Cetuximab. It is an inhibitor for the epidermal growth factor receptor and has been administered for treating metastased colorectal cancer [46] [47] and several types of cancer [48] [49] [50] [51]. The rationale behind using Cetuximab to treat TNBC patients is the expression of EGFR in 60% of this subtype [52] and the availability of this drug. However, treatment with Cetuximab in these patients showed low response rates in a clinical trial [53] [54].

The failure to repurpose an existing drug in this example means there is a pressing need for researchers to identify molecular makers for new drug developments. One example of how research can guide new drug development is the frequent identification of *BRCA1* mutations in TNBC patients. Hypermethylation of the *BRCA1* promoter is found to downregulate ER [55] and upregulate basal markers [56], thus leading to the proposal of its contribution to TNBC development. This suggests that poly Adenosine diphosphate (ADP)-ribose polymerase (PARP) inhibitors can possibly be used to treat these patients [42] [57].

A novel oncogene, *BCL11A* was recently identified by our group as a potential molecular marker that can address the challenge of improving prognosis for the TNBC subtype.

1.3 BCL11A

BCL11A protein is the human homologue of *EVI9* [58], originally found from studying myeloid leukaemia tumours in mouse models. This gene is highly conserved on chromosome 2 in humans, with 94% homology to mouse's *BCL11A* at the nucleotide level, 95% at the amino acid level [59], and 98% at the protein level [60]. Another member of the BCL11 family, B-cell Leukaemia 11B (*BCL11B*)

protein is 67% similar to *BCL11A* in terms of nucleotides and 61% at the protein level [60]. Moreover, they both bind to the consensus G-C rich motif at the promoter region of their target genes or in their binding partners presented in Table 1.2 [59] [61] [62].

BCL11A contains five exons with long introns encoding for Cys2His2 (C2H2)-like fold group type zinc-fingers protein. Alternative splicing of the exons gives rise to four isoforms: XS, S, L, and XL which differ in protein lengths and domains [59]. Figure 1.2 demonstrates the features and domains in these isoforms. All four of them encode for the zinc-finger at the N-terminus, along with the 12 conserved amino acids upstream. The main difference is the number of zinc-fingers at the C-terminus and thus the resulting protein length. Additionally, only L and XL isoforms have a proline-rich section and acidic domain from exon four [60]. Despite the similarity in size, Pulford et al. [63] successfully distinguished between the two isoforms and found the expression of XL to be the most abundant in cells.

1.3.1 Expressions and roles in biological systems

BCL11A is expressed diversely in tissues including the brain, kidney, haematopoietic stem cells (HSC) in the bone marrow, and other lymphoid tissues [59]. Isoforms also have preferences in spatial expressions; for example, the short form is expressed in erythrocytes progenitors [64] or foetal liver [58] while the long form is predominantly expressed in the bone marrow [59].

1.3.1.1 Roles in the immune system

1.3.1.1.1 Roles in silencing γ-globin chain and globin chain switching during development

BCL11A has a major role in regulating haemoglobin (Hb) F level [64]. Switching from γ- to β globin in Hb adult (HbA) is achieved by interacting *BCL11A* and its binding partners with one another and then binding them at various regions of β-globin gene cluster. These are shown in Table 1.2 and Figure 1.3. Dysregulation of

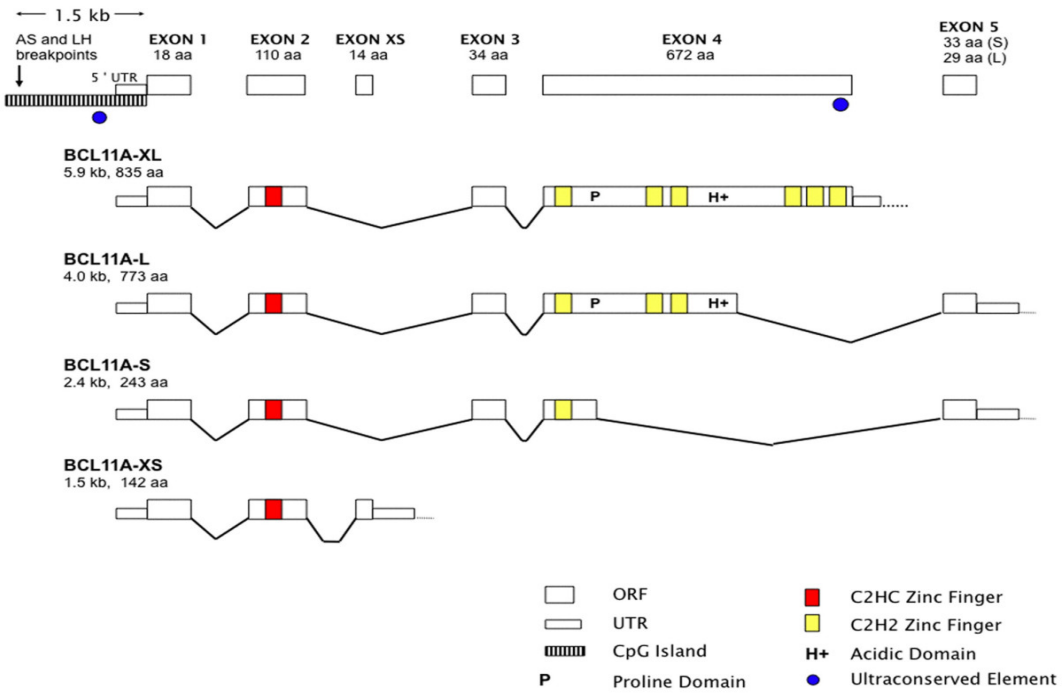


Figure 1.2: Structure of the *BCL11A* gene and the four isoforms.

The *BCL11A* gene has five exons with long introns in-between. It starts with a long CpG island with a highly conserved region, followed by the 5'untranslated region (UTR) prior to exon 1. An additional highly conserved region is present in the last zinc finger encoded on exon 4. Exon 2 translates to a zinc finger present in all isoforms and varying numbers of zinc fingers are translated from exon 4 in isoforms S, L, and XL. In addition to these isoforms, XS is created through alternative splicing. The conserved region among the CpG island, the 5' UTR downstream, and exons 1 and 2 are present in all isoforms. Exon 3 is present in S, L, and XL isoforms, whereas different fragments of exon 4 are found in isoforms S and L. XL is the only one with the full exon 4 which translates into six C2H2 zinc fingers. Moreover, a proline-rich domain is present between the first two zinc fingers from exon 4. [59]

Hb foetal (HbF) levels could result in sickle cell anaemia or β-thalassemia [62] [64] [65].

1.3.1.1.2 B-cells and T-cells developments

The second role of *BCL11A* in blood cells is to regulate the development of B- and T- lymphocytes during lymphopoiesis [66]. B-cell differentiation and cell survival are regulated by haematopoietic regulator BCL6 [64] and anti-apoptotic BCL2 protein, respectively. At the same time, *BCL11A* regulates T-cell maturation by utilising the Mdm2-p53 pathway [68]. Mutations could lead to B-cell lymphoma leukaemia by transrepressing genes through histone deactylase, SIRT1, and to T-cell lymphoma by triggering over-production through Notch 1 signalling [69].

1.3.1.2 Roles in the neural system

BCL11A has also been identified to have a number of roles in the neural system, such as axon branching and dendrite outgrowth. These are achieved through the guidance by regulating Class 3 semaphorins to interact with the plexin transmembrane receptors [71]. Outside of participating in different functions by regulating genes and proteins in the neural system, *BCL11A* when mutated is also found to be the cause of various neuronal diseases. For example, microdeletion of *BCL11A* results in autism spectrum disorder (ASD) [71] and single nucleotide polymorphism (SNP) in Schizophrenia [72].

1.3.1.3 Roles in Lung cancer

Apart from the identification of *BCL11A* in TNBC, studies by our group have also found *BCL11A* to be an oncogene in Lung squamous cell carcinoma (LUSC) [73]. LUSC is one of the non-small cell lung cancers (NSCLC) with limited targeted therapies and patients are often treated with Platinum-based drugs as the first line of defence [73]. Our results have shown *BCL11A* to regulate transcription factor *SOX2*, and the two regulate *SETD8* together. Inhibiting *SETD8* also resulted in the inhibition of tumour growth which suggests that *BCL11A-SOX2* interaction has potential as a new drug target.

Binding partners	Functions
SOX6	Directly interacts with BCL11A to regulate globin switching [65]
Matrin-3	Assists BCL11A binding to nuclear matrix and interacts with β-globin gene [63]
GATA 1 – FOG 1	Bridging between BCL11A and distant site of β-globin gene to regulate globin switching [63] [70]
NuRD complex	Anti-apoptotic protein of B- and T-cells survival [63]
BCL2	Hematopoietic regulators for germinal centre formation [68]
BCL6	Member of Histone Deactylase family regulating transcription repression [62] [67] [68]
SWI/SNF complex	Chromatin remodelling for transcription machineries to access and transcribe genes [65]

Table 1.2: Examples of *BCL11A* binding partners and their roles in relation to the functions of *BCL11A*.

The various roles of the binding partners of *BCL11A* are presented above.

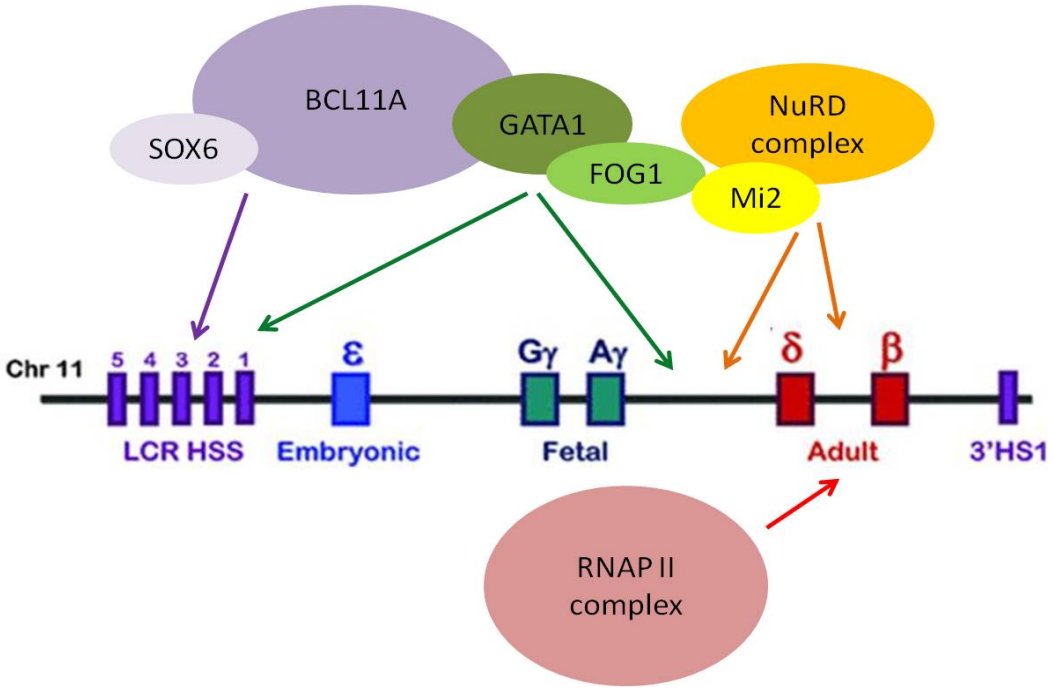


Figure 1.3: Structure of the β- globin gene clusters.

The β- globin gene clusters are found on chromosome 11, with Locus control region (LCR) at the 5'end and Haematopoietic lineage cell-specific protein 1 (HS1) at the 3'end. It also has multiple globin genes expressed in a developmentally sequential order - Embryonic (ε), Fetal (G gamma (γ) and Aγ), Delta (δ) and Beta (β). BCL11A works in concert with multiple proteins to form a complex in silencing fetal γ gene and transcribing adult β gene during development. These are part of the subunits which form tetrameric Hb. LCR is a common site bound by BCL11A and other proteins such as SOX6 and GATA1. Haematopoietic regulators such as GATA1 and FOG1 bind to the genomic region between γ and δ, a site that the Mi2/NuRD repressor complex also binds to in addition to the gap between δ and β. All these proteins work as one to silence γ-globin and initiate the transcription of β-globin by recruiting RNA polymerase (RNAP) II complex. (Modified from [64] [65])

1.3.1.4 Roles in Triple-negative breast cancer

I have described the regulation of *BCL11A* in different biological systems in the above sections, but its regulation in the context of TNBC is not yet known. Attempts have been made by our group to understand this. It has been extensively demonstrated that *BCL11A* regulates the development of haematopoietic cells, and mutations can lead to malignancies. Considering the regulations in B-cell lymphoma leukaemia, one would speculate *BCL11A* to transrepress genes in a similar manner through histone deacetylase SIRT1. Nevertheless, the level of SIRT1 remains unchanged in mouse models where carcinogen 7,12-Dimenthylbenz[a]anthracene (DMBA) was used to induce tumourigenesis [74]. This raises the possibility that *BCL11A*'s regulation is context dependent and thus it is important to investigate the roles and regulations of *BCL11A* in the TNBC subtype.

Screening from a clinical dataset has found *BCL11A* to be highly expressed and with genome instability, as demonstrated by an increase in copy number and gene expression in TNBC. *BCL11A* was also found to have hypomethylated specifically in these tumour cells [74]. Hypomethylation can be a de novo or heritable epigenetic change and is often found in satellite DNA or oncogenes [75]. Demethylation is a process which removes methyl groups from cytosine and adenine bases, allowing transcription machineries to access or recruit histone associated proteins.

Experimental works have been conducted to investigate the role of *BCL11A* in the context of TNBC. Increase in mammospheres and tumour sizes were observed by overexpressing *BCL11A in vitro* and *in vivo*, respectively; knocking-down its expression has also led to a reduction in tumour size. These suggest its importance in tumourigenesis. Reduction in tumour sizes through *BCL11A* deletion suggests its involvement in tumour maintenance. All these evidence support the role of *BCL11A* as an oncogene of TNBC. Screening through clinical databases further demonstrates the specificity of *BCL11A* expression found in this particular subtype.

However, the regulation of *BCL11A* expression and downstream pathways in Triple-negatives awaits to be discovered.

1.4 Transcriptional regulation

Before looking deeper into the regulation of *BCL11A*, I will first give a general overview of transcription and examples of regulating this process. Transcription is important for regulating gene expression and transcribing deoxyribonucleic acid (DNA) into messenger ribonuclease acid (mRNA) by ribonuclease acid (RNA) polymerase for transferring genetic information. These mRNA will then be translated into proteins. Figure 1.4 presents the major steps in transcription: initiation, elongation, and termination. Transcription starts at the initiation stage by recruiting a pre-initiating complex, composed of many transcription factors with general or specific roles, to a transcription start site and initiates by adding a complementary nucleotide. The RNA polymerase then moves along the DNA strand while extending the mRNA, in a process called elongation. Finally, termination occurs upon the arrival of the RNA polymerase at a stop codon that signals it to terminate transcription, release the mRNA transcript, and dissociate from the DNA strand.

There are a number of ways to regulate transcription. For example, epigenetic control modifies DNA or histone proteins for the accessibility of DNA sites for transcription [76]. This can result in gene activation or repression. An alternative mechanism involves DNA forming a loop between the promoter regions and enhancers or repressors, followed by the recruitment of transcription factors [77].

1.4.1 Transcription factors

Transcription factors are proteins which bind at specific DNA sequences to regulate gene expression. This tight regulation is important for many cellular processes such as cell growth, differentiation, and survival. Transcription factors can act alone or in concert with other proteins to activate or repress genes by modulating the

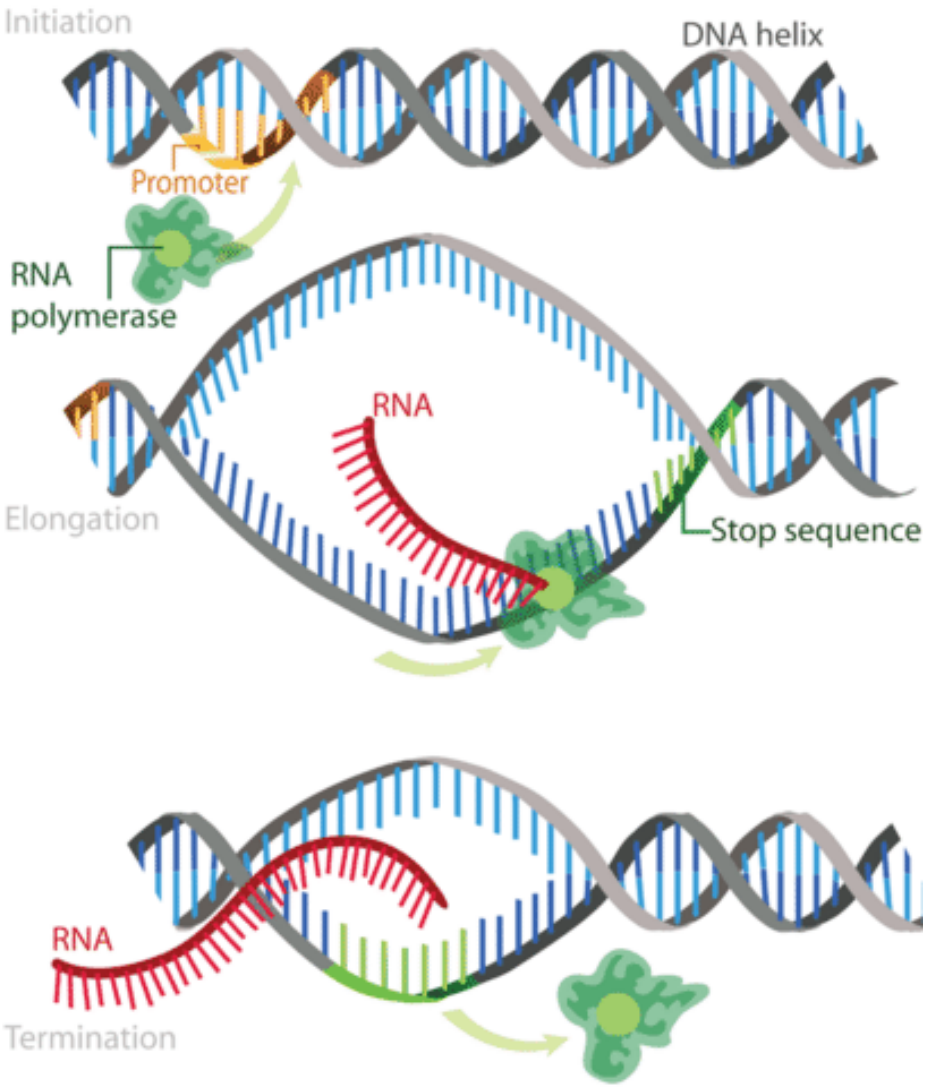


Figure 1.4: An overview of the processes in transcription.

Initiation starts when RNA polymerase binds close to the promoter region on the DNA and recruits a pre-initiating complex. RNAP then moves along the DNA strand while transcribing and extending the new mRNA transcript. This process is called elongation. Termination occurs when RNAP reaches the sequence that code for a stop codon. The RNAP then releases the mRNA transcript and dissociates from the DNA strand. (Obtained from [78])

accessibility of DNA regions needed for transcription, or by recruiting co-activators and co-repressors.

It is known that tumour cells have upregulated levels of transcription and targeting cell proliferation has been a common thought for treating cancer patients. Proteins such as transcription factor are one type of possible targets. Proteomics is the name given to the study of protein, And in this project, I will be using the proteomics approach to study transcriptional regulations of *BCL11A*.

1.5 Proteomics studies

Advancement in technology has given rise to new ways for classifying breast cancer patients, one of which is gene expression profiling. This allows further characterisations and tailored treatments based on a patient's genetic component.

However, the use of gene expression has a disadvantage in that transcripts are not necessarily translated into proteins. Moreover, it is estimated that the human genome contains 20,000 protein-coding genes which can give rise to 500,000 protein isoforms through alternative splicing and post-translational modifications [79]. As a result, genetic information alone may not be detailed enough for tailoring specific treatments to individual patients. Hence, the study of proteomics can add another level of accuracy and sophistication.

Another reason for the study of proteomics is that it allows us to relate between genome and protein functions. In particular, genome is related to both the expression and abundance of proteins. Therefore, understanding the relationship between genome and protein functions can aid us in disease prediction. In close connection to this, the study of proteins may also lead to the development of new biomarkers for early detections and improved prognosis.

Yet another reason for studying proteomics comes from the fact that a protein can potentially participate in a large number of roles; see for instance the earlier

example of *BCL11A* in different biological systems. By understanding these different roles of proteins, we can predict their involvement in more complex networks in cancer, for instance in signalling pathways, tumour progression, and response to therapies.

In summary, proteomics can help to predict diseases, develop new therapeutic ways, and overcome clinical challenges when treating breast cancer patients.

1.5.1 Traditional methods to study proteomics and Rapid Immunoprecipitation of endogenous proteins with mass spectrometry (RIME)

For decades, protein identification has been achieved through two-dimensional gel electrophoresis (2D-PAGE) [80]. This technique first separates proteins on a gel based on their isoelectric point and then by their molecular weight, thus producing a specific 'dot' on the gel. Each of these 'dots' can then be passed on to mass spectrometry, a widely used technique in proteomic studies, for protein identification.

The identification of protein leads to the study of protein-protein interactions using various affinity chromatography techniques. Potential candidates are then validated using Fluorescence resonance energy transfer (FRET) and Proximity ligation assay (PLA) [81]; both present results as signals when protein partners are in close proximity. Alternatively, they can be validated using co-immunoprecipitatioin (Co-IP) which pulls-down partners using one as bait and vice versa. However, Co-IP has several drawbacks. Multiple washings between steps could result in the loss of proteins with transient interactions. Previous knowledge of the specific protein-protein interactions is required before Co-IP can be performed. Poor quality antibodies may also produce false negatives [82]. Finally, endogenous proteins are found at a lower concentration in their native state than when used experimentally, and this difference in expression levels may also explain the failure in validating some of the protein partners experimentally.

To overcome these problems, RIME is employed. RIME is a technique developed recently to study and identifies unknown proteins. It couples chromatin immunoprecipitation (ChIP) with mass spectrometry to identify DNA-protein or protein-protein interactions in living cells [78]. ChIP is a commonly used technique to study DNA-protein interactions in cells, and its adaptation into the RIME protocol changes the application of ChIP from studying interactions to pulling-down proteins or partners on a specific piece of DNA for later identification using mass spectrometry. Figure 1.5 outlines the steps in RIME.

Similar to ChIP, cells are first fixed using formaldehyde to fixate and preserve transient or long-lasting, weak or strong interactions [83]. Formaldehyde is commonly used in ChIP and cell fixation protocols due to its small size, non-specificity to targets, reversibility, and the ability to cross membranes [83]. To study transcription factors, the nuclear fractions are then extracted from the cell pellets and broken up into small DNA fragments. They are then immunoprecipitated to purify the protein of interest with its binding partners and further processed with mass spectrometry to identify the unknown proteins. The advantage of using RIME here is that no prior knowledge of potential protein partners is needed, whereas traditional methods such as affinity chromatography may require some knowledge such as the size or charge of the unknown protein [84]. The RIME technique is used in this project to study transcription factors which bind at specific *BCL11A* genomic regions and regulate its expression.

1.5.2 Applications of RIME

Since its development, RIME has been used in a number of studies. For example, protein partners were identified for maintaining stemness in the human trophoblast stem cells for controlling placenta development [85]; for tumourigenic functions in prostate cancer [86]; and as part of a complex to function as a tumour suppressor [87].

A good example of using RIME to identify transcription factors is the identification of Growth regulating Estrogen receptor binding 1 (GREB1) in ER-positive breast

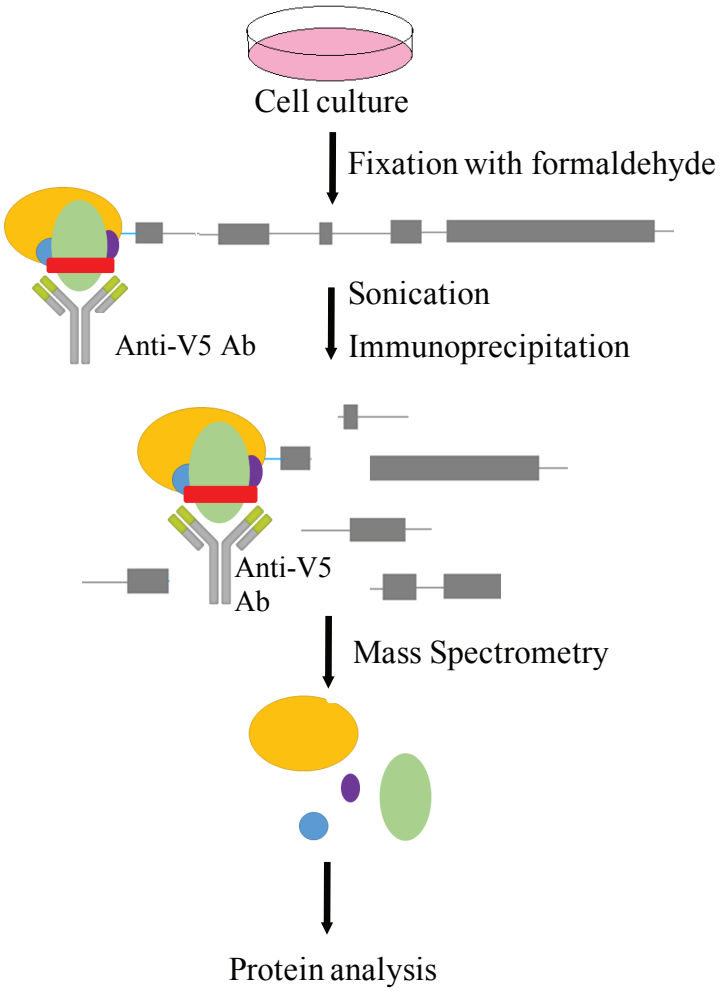


Figure 1.5: The major steps in RIME.

Cell cultures are fixed with formaldehyde. Diagrams onwards represent actions at the genomic level. Primers from the CRISPR-Cas9 system are bound to the target sequences through complementary base pairing with the help of catalytically inactive Cas9. Nuclease activities of Cas9 are silenced, allowing the RNA-hybrid Cas9 complex to bind to target DNA without inducing double-strand breaks. Sonication breaks up the loci into small DNA fragments and V5-bound antibodies are used in immunoprecipitation to pull down multi-tagged dCas9 along with any proteins bound in close proximity during the fixation stage. Samples are then processed by mass spectrometry to identify possible protein partners with RIME.

cell lines [88]. It has also been demonstrated that through coupling RIME with Stable isotope labelling by amino acids (SILAC) in cell cultures, investigators could identify differences in protein expression profiles under the effect of a drug [88]. Another example is the identification of Chromodomain helicase DNA binding protein 8 (CHD8) interacting with BCL11A in TNBC by our group.

1.6 Genetic engineering and genome editing

In addition to using RIME to identify putative transcription factors which may regulate *BCL11A* in TNBC, I also employ the CRISPR-Cas9 system for targeting specific regions of *BCL11A*.

1.6.1 The CRISPR-Cas9 system

1.6.1.1 Mechanism of the system

CRISPR was originally used as a Type II adaptive immunity against foreign genetic material in bacteria or archea [89] [90]. This locus contributes to 45% of the bacterial genome [91].

Foreign DNA such as viral DNA or plasmids contains short sequences of 20 nucleotides called spacers. Upon first encounter, these spacers are cleaved and incorporated into the host's DNA between the repeat sequences forming CRISPR. Bacteria and archea avoid secondary infections through a RNA-guided DNA recognition mechanism called Cas system, as illustrated in Figure 1.6. The Cas system starts with transcribing the CRISPR locus to pre-crRNA which is then processed into crRNA containing partial repeats and particular spacer sequences. This maturation process requires additional RNAs and proteins. Trans-encoded small RNA (TracrRNA) recognises the repeat in the pre-crNRA sequence and forms a hybrid with it. Excess sequences are cleaved off by host's RNase III and forms a RNA-protein complex with endonuclease Cas9 [89]. CRISPR-Cas9 system identifies target DNA through complementary base pairings between the 'seed sequence' at the 5' end of crRNA and target sequences [92]. Target recognition is

further secured by the presence of protospacer-adjacent motif (PAM) immediately downstream of the spacer [93]. This PAM site is important for multiple reasons. It is crucial for Cas9 binding as each ortholog has its own specific sequences. For example, *Streptococcus pyogenes* Cas9 has sequence NGG at its PAM site [94]. It is also important for distinguishing between self and foreign target sequences. Subsequent to DNA binding, Cas9 plays an important role in inducing double-strand breaks with its HNH and RuvC nuclease domains [95]. These breaks are repaired by two DNA damage mechanisms described in Figure 1.7: the error prone Non-homologous end joining (NHEJ) that has a higher chance of introducing indel and frameshift mutations, and an alternative mechanism through precise homology directed repair (HDR) but which requires a template [96].

1.6.1.2 Comparison between traditional genome editing techniques and CRISPR-Cas9

Traditional genome editing approaches such as Zinc finger nucleases (ZFNs) and transcription-activator-like effector nucleases (TALENs) both rely on protein-DNA recognitions and require the design of a new nuclease for every target [99] [100]. These systems works by fusing the DNA-binding domains with endonucleases which induce double-strand breaks for genome engineering. In comparison to these techniques, the CRISPR-Cas9 system offers more simplicity in its design and application for genome engineering. Instead of having to design a new nuclease per target, users design a single guide RNA (sgRNA) for their specific targets. sgRNA is a fusion of crRNA and tracrRNA in the natural system and this alone is enough to guide the RNA-recognition system to the target DNA sites [101].

1.6.1.3 Designing components for the CRISPR-Cas9 system

The first part of refining this CRISPR-Cas9 technique is to custom design a gRNA which is complementary to the site of interest. This gRNA sequence is equivalent to the crRNA in the natural system for guiding the system to specific targets. Thus, it is important to optimise the design of gRNA for effective targeting while minimising off-target effects. A critical requirement includes having a PAM sequence immediately following the gRNA at the 3' end for Cas9 recognition [92]

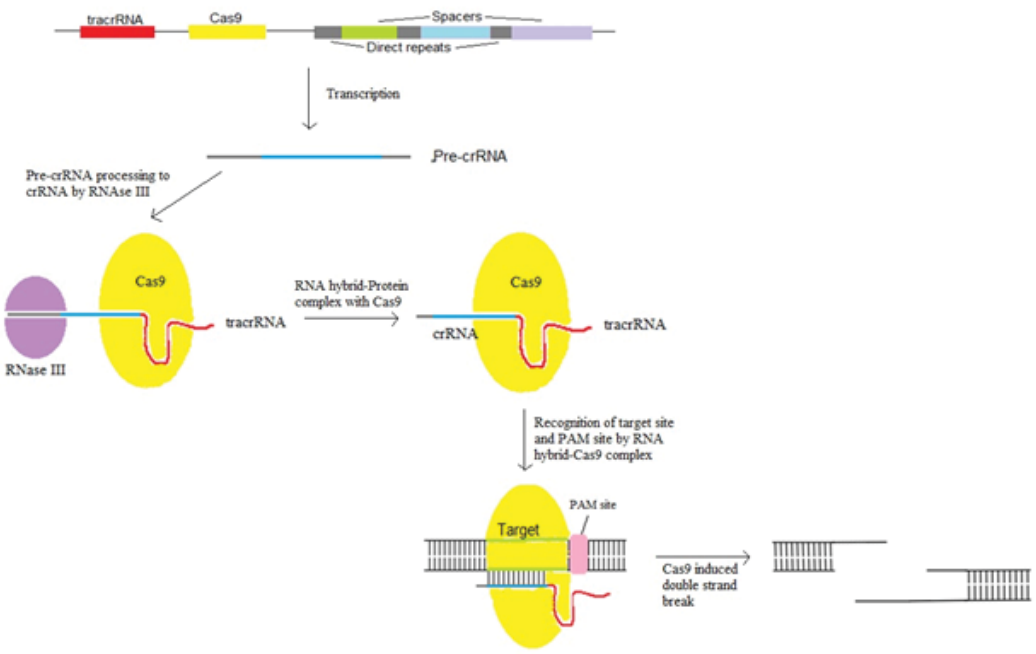


Figure 1.6: The CRISPR-Cas9 system.

The CRISPR locus consists of tracrRNA, Cas9 and spacers separated by repeats. Each spacers are unique and acquired from each invasion by foreign genetic materials in bacteria and archea genome. Upon second invasion, the corresponding spacer would be transcribed with parts of the repeat sequences to form pre-crRNA. This immature transcript is recognised by tracrRNA and endonuclease Cas9. Additional protein from the host RNase III is also required to cleave off excess repeat sequences to form mature RNA hydrid-Cas9 complex. crRNA is complementary to the target sequence and this brings the complex to the target. This system is also called RNA-guided DNA recognition. The PAM site immediately following the target is also critical for Cas9 recognition and binding. Cas9 is an endonuclease with two catalytic domains: HNH and RuvC domains which induce double-strand breaks and stopped the invasion.

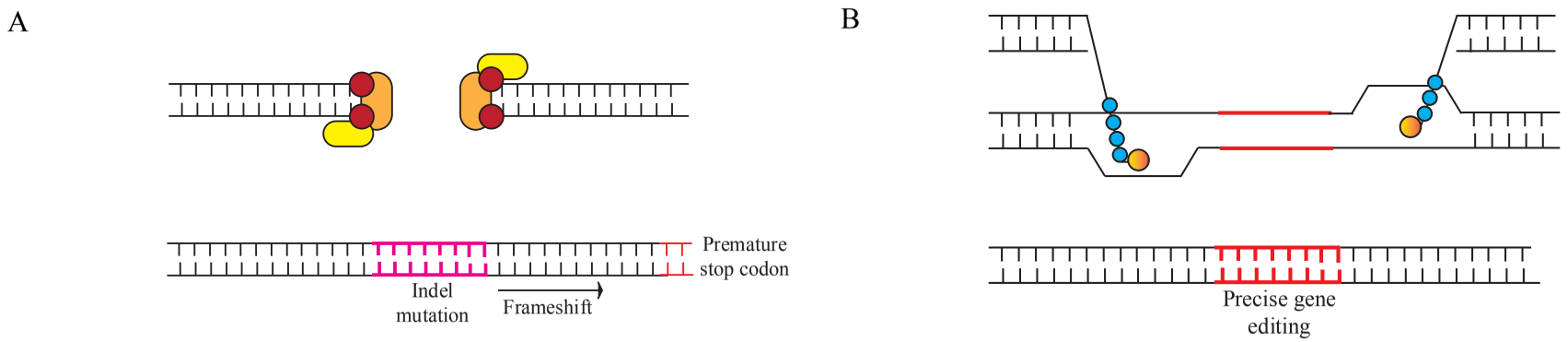


Figure 1.7: The two pathways to repair double-strand breaks.

The mechanism of Non-homologous end joining is shown in (A). Double strand breaks are first detected and bound by protein complexes including Ku heterodimer and DNA-PKcs. Artemis is one of the DNA-Kcs that can trim the mismatch nucleotides through its nuclease activity. Strands are then re-synthesis by DNA polmyerases. The final product after the ligation by DNA ligase can result in indel mutations which are part of the frameshift mutations that introduce premature stop for transcription. (B) shows the mechanism for Homology directed repair. 3’ overhangs are created at the double stand breaks. These single stranded regions are bound by RPA and Rad51 proteins. Rad51 then catalyses strand transfer between the broken strands and the homologous templates by forming a D-loop. This results in DNA recombination and precise gene editing. [97] [98]

[93]. This PAM sequence is specific to the species of Cas9 used and could potentially affects the positions and activities of the gRNAs. The specificity of gRNA and the number of similar sites in the genome also contribute to potential off-targeting. The best site that matches these requirements can be predicted using publicly available web-based tools that allow gRNA to be designed based on the genomic regions of interest and built-in computational algorithms that predict their activities [102].

1.6.1.4 Applications of the CRISPR-Cas9 system

Cas9 is an effective nuclease which can be used to delete or correct DNA mismatches. The CRISPR-Cas9 system can also be employed in genetic engineering for introducing gene knock-out or knock-in [103], generating transgenic animal models [104], modulating transcriptional regulations [105], epigenetic modifications [106] or, in the case of this project, be used as a tag in a novel way.

DNA double strand breaks induced by Cas9 can be repaired through the error prone NHEJ repair pathway. This is employed in gene knock-out experiments. Repair through this pathway introduces indel mutations which result in frameshift mutations [107] and premature stop of target gene's transcription. On the other hand, DNA repair through the HDR pathway is utilised in gene knock-in experiments by providing template for the homologous recombination [108] and introducing the desired gene into cells or animal models. This HDR pathway was used in our group to generate *BCL11A* transcriptional reporter lines for studying genes which may regulate *BCL11A* in TNBC. Our results have identified E2F transcription factor 4 (*E2f4*) and SET domain-containing protein 4 (*Setd4*) as potential novel regulators of *BCL11A*.

In addition to the use of the DNA repair pathways, multiple forms of Cas9 can also be employed for various objectives for genetic engineering. An example would be to include single mutations in one of the Cas9's nuclease domain and turn it into a 'nickase'. DNA binding could also be uncoupled with the nuclease activity, by

mutating H480A and D10A in HNH and RuvC doamins respectively in catalytically inactivate Cas9 (dCas9) [109]. Its mechanism is demonstrated in Figure 1.8.

dCas9 can also be employed to regulate transcriptional activities by fusing with an effector, such as an activator or repressor, and guide it to sites close to the promoter to control transcription or affect epigenetic control [110]. However, one disadvantage of this system is Cas9 itself may affects gene expression by blocking the genetic regions and accessibilities of DNA for transcriptional machineries; this is referred to as CRISPR interference (CRISPRi).

1.7 RNA interference technology

RNA interference (RNAi) technology has been frequently used for gene silencing. It was originally discovered in *C.* elegans by Fire and Montagomery [111]. Injection of double stranded RNA (dsRNA) has led to the degradation of mRNA which has the same sequence as these dsRNA [111]. RNAi is now a technique employed to inhibit gene expression or translation using exogeneous RNA molecules. One of these is the small interfering RNA (siRNA) that binds to the target mRNA transcript in a sequence-specific manner, thus creating dsRNA which leads to the degradation of mRNA and downregulation at the protein level [112].

However, one disadvantage for siRNA is the short term expression and degradation by cellular machineries [113]. These can be overcome by the use of short hairpin RNA (shRNA) which combines siRNA with a loop structure called the T-loop and allows constitutive expression in the cell [113]. Figure 1.9 presents the mechanism of siRNA and shRNA. shRNA are first delivered into the cells through transfection [114] or viral transduction [115]. They are transcribed in the nucleus by RNA polymerase III called Drosha, transported into the nucleus and then modified by RNase III Dicer to remove the T-loop and create siRNA with 20-25 nucleotides and a 3’ overhang. This siRNA is then loaded onto a RNA-induced silencing complex (RISC) with Argonaute family of proteins. Next, the double stranded siRNA in the

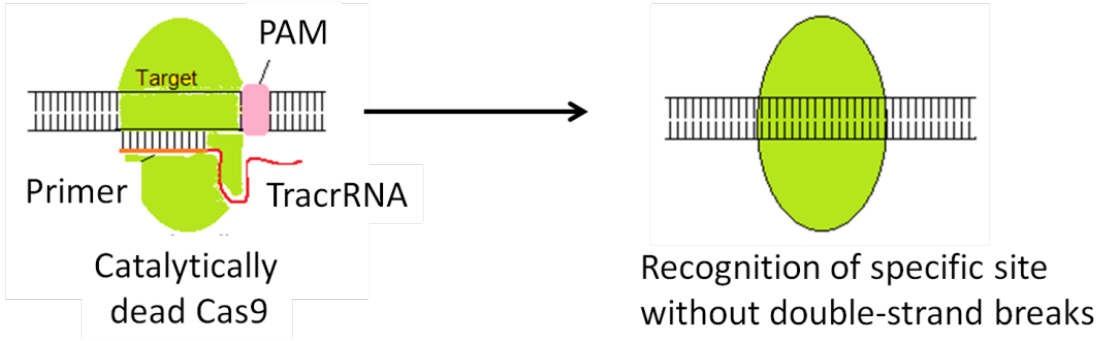


Figure 1.8: The mechanism adapted by catalytically inactive Cas9 (dCas9).

Same requirements as figure 1.6: The presence of PAM site immediately downstream of target sequence and complementary base pairings between target and specifically designed primer. The only differences are the inactivation of HNH and RuvC domains which allow the target to be recognised but not cleaved.

RISC unwinds and dissociates, leaving the sense strand as a guide for target recognition through complementary base pairing. Upon target identification, the RISC complex cleaves the target RNA's phosphate backbone. The siRNA-target mRNA double stranded RNA duplex is then dissociated from the RISC complex and degraded by cellular machineries [116].

This RNAi technology is used in this project post-identification of protein regulators of *BCL11A* for validation and functional investigations.

1.8 Aims and outlines of this project

The main aim of this project is to investigate the regulation of *BCL11A*'s expression in TNBC. To achieve this, I will use the CRIPSR-Cas9 system in combination with proteomics. The CRISPR-Cas9 system was used to target specific sequences from the *BCL11A* genome and RIME ultimately to identify proteins regulating *BCL11A*'s regulation in the TNBC subtype.

Chapter 3 describes the experimental design of incorporating the CRISPR-Cas9 system into our studies. In this project, I will use the catalytically dead form of Cas9 to study the transcription regulation of *BCL11A*, focusing on regions which display active transcriptions. Complementary sequences were then designed using the CRISPR design tool. These target complementary primers act as crRNA in the CRISPR-Cas9 system for the RNA-guided DNA recognition. Along with the Cas9-expressing plasmid, primers were cloned into a different expression vector and delivered into MDA-MB-231 TNBC cell line. Upon establishing stable cell lines, specific antibodies for Cas9 would be used to pull down dCas9 with proteins sitting at close proximity and identified using the RIME proteomic technique.

Chapter 4 describes the steps and rationale for interpreting the RIME dataset using multiple parameters to identify potential candidates regulating *BCL11A*. These include selecting proteins with nuclear expressions; selecting proteins against a database which identifies common contaminants in proteomic studies and against

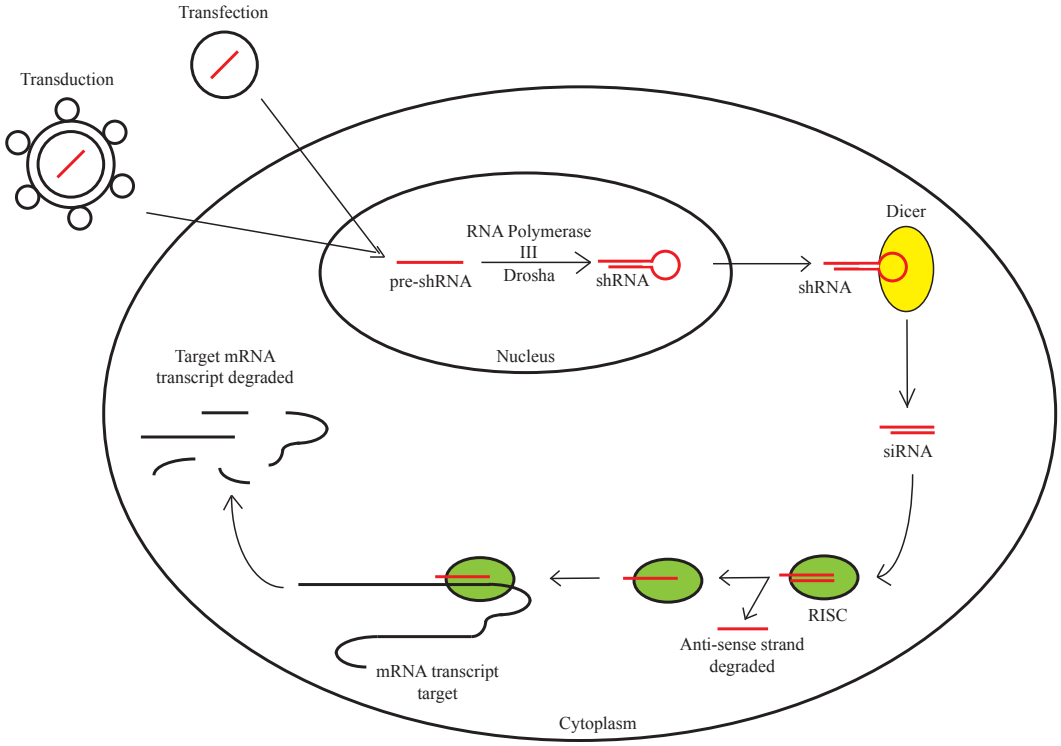


Figure 1.9: The mechanism of the RNA interference technique.

shRNA can be delivered into the cells by transfection or viral transduction. The pre-shRNA transcript is transcribed by RNA polymerase III called Drosha into mature shRNA. It is then transported into the cytoplasm and processed by RNase III Dicer into siRNA. The siRNA is then loaded onto RISC, unwinding of the RNA duplex result in the anti-sense strand degraded by cellular machineries and the sense strand to recognising the target mRNA transcript by complementary base pairing. Upon recognition, RISC then cleaves the target followed by degradation by cellular machineries and a reduction in level of protein translated. [112] [116]

the cancer database to identify proteins which upregulation in TNBC and relations with *BCL11A*. These expressions would then be validated using western blot on a panel of breast cancer cell lines; DNA-protein interactions using ChIP and protein-protein interactions by Co-immunoprecipitation.

Chapter 5 describes the investigation on the effects of these potential regulators on the impact of *BCL11A*'s expression and functional effects using shRNA.

CHAPTER 2

Materials and Methods

2.1 Methods

2.1.1 Cloning

2.1.1.1 Guide RNA design and cloning into PiggyBac vector

gRNA sequences targeting specific sites on the human *BCL11A* gene locus chr2: 60,684,329-60,780,633 were designed by computational screening using the University of California – Santa Cruz (UCSC) genome database browser [117], named Sites A and B from now on, and ordered in powder form from Sigma. These sequences shown in Table 2.1 were selected based on the following criteria: 1) H3K27Ac histone mark and 2) hypersensitivity to DNaseI cleavage. The forward gRNA primers had CTTG attached at the 5' end and complementary sequences with AAAC at the 3' end. 1M of each primer was mixed into a 10 µl reaction, incubated at 95^{o}C for 5 minutes followed by rampdown at a rate of 0.2^{o}C per second to room temperature to form oligos.

PiggyBac vector, PB-gRNA-BsaI-EF1a-RCB (courtesy of Pentau Liu, Sanger), was digested in a 20µl reaction containing 2µl of restriction enzyme BsaI-HF (NEB) and 2µl of cutsmart buffer (NEB) at 37^{o}C for 3 hours followed by gel electrophoresis. DNA fragments with 9000 base pairs were excised and eluted using QIAuick Gel Extraction Kit (Qiagen). 1:1 digested gRNA expression vector and gRNA oligos, 1µl of T4 DNA ligase (NEB) and 1µl of T4 DNA ligase buffer (NEB) were mixed to form a 10µl reaction mixture and incubated at room temperature for 2 hours. 5µl from the ligation reaction were used to transform NEB 10-beta Competent Escherichia coli (*E.coli*) (NEB). Colonies were picked and 5-10µl of bacterial culture was extracted using QIAprep Spin Miniprep Kit (Qiagen). Oligo insertions were then confirmed by double digestions using restriction enzymes NotI-HF (NEB) and HpaI (NEB).

Primers	Sequences
Site A forward	CTTG CGGTGGTGAGATGACCGCCT CGG
Site A complementary	AAAC CCGAGGCGGTCATCTCACCACCG
Site B forward	CTTG GAGAGGTCCACGAGCCACGC AGG
Site B complementary	AAAC CCTGCGTGGCTCGTGGACCTCTC

Table 2.1: The gRNA sequences designed for targeting *BCL11A* with dCas9

The sequences used as gRNAs against Sites A and B for the plasmid, PB-gRNA-BsaI-EF1a-RCB, are shown with CTTG at the 5' and AAAC at the 3' ends followed by a PAM sequence.

2.1.1.2 Knocking-out *BCL11A* with CRISPR-Cas9 vectors

Primers targeting Sites A and B of human *BCL11A* were designed by computational screening using the UCSC genome database browser and ordered in powder form from Sigma. The forward gRNA primers had CACC attached at the 5' end and complementary sequences with AAAC at the 3' end. These primer sequences are shown in Table 2.2. 5µl of forward and complementary primers were annealed together at 95ºC for 3 minutes followed by rampdown to 22 ºC at the rate of 0.1ºC per second.

The CRISPR-Cas9 vectors px458 and px459, pSpCas9(BB)-2A-GFP (Addgene) and pSpCas9(BB)-2A-Puro (Addgene), respectively, were digested in a 20µl reaction with 2µl of BbsI (NEB) and 2µl of Buffer 1.1 (NEB) at 37ºC for 1.5 hours. DNA fragments with 9000 base pairs were excised after gel electrophoresis and extracted using QIAuick Gel Extraction Kit. Next, 4µl of digested vector were incubated with 4µl of annealed oligo, 1µl of T4 DNA ligase (NEB) and 1µl of T4 DNA ligase buffer at room temperature for 2 hours. 5µl from the ligation reaction were used to transform NEB 10-beta competent cells. Colonies were picked and 5-10µl of bacterial culture was extracted using QIAprep Spin Miniprep kit. Positive clones were checked by double enzyme digestions using BbsI (NEB) and AgeI (NEB).

2.1.1.3 Designing shRNA primers and cloning into PiggyBac vector

shRNA primers in Table 2.3 were obtained from MISSION® shRNA (Sigma) (https://www.sigmaaldrich.com/life-science/functional-genomics-andrnai/shrna/individual-genes.html) [118] and ordered in powder form from Sigma. 0.1M of each primer was mixed into a 20µl reaction with 2µl of T4 ligase buffer (NEB) and 1µl of T4 PNK enzyme (NEB), incubated at 37ºC for 1 hour for phosphorylation and then inactivated at 65ºC for 20 minutes. 5µl of each complementary primer pairs were then annealed together at 95ºC for 3 minutes followed by rampdown to 22ºC at the rate of 0.1ºC per second.

The Piggybac shRNA vector, PB-PGK-GFP-Neo-HI (Modified and used by our

Primers	Sequences
Site A gRNA 1 forward	CACC GTCTCCCTCCCACAAACTGG
Site A gRNA 1 complementary	AAAC CCAGTTTGTGGGAGGGAGAC
Site A gRNA 2 forward	CACC GAGCAGCGGGGAGACCACGG
Site A gRNA 2 complementary	AAAC CCGTGGTCTCCCCGCTGCTC
Site A gRNA 3 forward	CACC GGGGCTTTTACTTCGGCCCC
Site A gRNA 3 complementary	AAAC GGGGCCGAAGTAAAAGCCCC
Site B gRNA 1 forward	CACC GGCTTACAGATGACGCTCTG
Site B gRNA 1 complementary	AAAC CAGAGCGTCATCTGTAAGCC
Site B gRNA 2 forward	CACC GGGGTTCCAAACCAGGGCAG
Site B gRNA 2 complementary	AAAC CTGCCCTGGTTTGGAACCC
Site B gRNA 3 forward	CACC GTAAAACCCTTTCACTCATA
Site B gRNA 3 complementary	AAAC TATGAGTGAAAGGGTTTTAC

Table 2.2: The gRNA sequences designed for knocking out Sites A and B.

The forward and complementary sequences used in the CRISPR-Cas9 vectors, px458 and px459, for knocking-out corresponding regions of Sites A and B. Forward primer was attached with CACC at the 5' and complementary with AAAC at the 3' end.

Primers	TRC Number	Sequences
Scrambled sequence forward		GATCCCTAAGGTTAAGTCGCCCTCGCTC GAGCGAGGGCGACTTAACCTTAGG
Scrambled sequence reverse		TCGACCTAAGGTTAAGTCGCCCTCGCTC GAGCGAGGGCGACTTAACCTTAGG
Human *DEK* shRNA 1 forward	TRCN0000235740	GATCCCGGTGAAATTGAGAGGATACATTT CTCGAGAAATGTATCCTCTCAATTTCATTT TTG
Human *DEK* shRNA 1 complementary		TCGACAAAAATGAAATTGAGAGGATACAT TTCTCGAGAAATGTATCCTCTCAATTTCAC CGG
Human *DEK* shRNA 2 forward	TRCN0000235736	GATCCCGGCCCTACAGATGAAGAGTTAAA CTCGAGTTTAACTCTTCATCTGTAGGGTTT TTG
Human *DEK* shRNA 2 complementary		TCGACAAAAACCCTACAGATGAAGAGTTA AACTCGAGTTTAACTCTTCATCTGTAGGGC CGG
Human *SSRP1* shRNA 1 forward	TRCN0000343894	GATCCCGGCGCTTCGATGAGATCTCCTTTC TCGAGAAAGGAGATCTCATCGAAGCGTTT TT
Human *SSRP1* shRNA 1 complementary		TCGAAAAAACGCTTCGATGAGATCTCCTTT CTCGAGAAAGGAGATCTCATCGAAGCGCC GG
Human *PSIP1* shRNA 1 forward	TRCN0000074819	GATCCCGGGCAGCAACTAAACAATCAAAT CTCGAGATTTGATTGTTTAGTTGCTGCTTT TTG
Human *PSIP1* shRNA 1 complementary		TCGACAAAAAGCAGCAACTAAACAATCAA ATCTCGAGATTTGATTGTTTAGTTGCTGCC CGG
Human *PSIP1* shRNA 2 forward	TRCN0000298567	GATCCCGGTTTAGGACCAAAGGATATATT CTCGAGAATATATCCTTTGGTCCTAAATTT TTG
Human *PSIP1* gRNA 2 complementary		TCGACAAAAATTTAGGACCAAAGGATATA TTCTCGAGAATATATCCTTTGGTCCTAAAC CGG

Table 2.3: The shRNA sequences obtained for candidates knock-down

gRNA sequences for knocking-down the three proteins were obtained from Sigma and the scrambled sequences as a control.

group, refers to [74]), was digested in a 20µl reaction with 2µl of Xho1 (NEB), 2µl of BglII (NEB) and 3µl of Buffer 3.1 (NEB) at 37ºC for 2 hours followed by the addition of 1µl of CIP (NEB) and incubated at 37ºC for 30 minutes. Reactions were then run on agarose gel, after which DNA fragments with 6000 base pairs were cut out and extracted using QIAuick Gel Extraction Kit. 0.1µg of digested vectors was ligated with 6.5µl of annealed primers, 1µl of T4 ligase (NEB) and 1µl of T4 ligase buffer (NEB) in a 10µl reaction and incubated at room temperature for 2 hours. 5µl from the ligation reaction were used to transform NEB 5α Competent *E.coli* (NEB). Colonies were picked and 5-10µl of bacterial culture was extracted using QIAprep Spin Miniprep kit. Oligo insertions were determined by double digestions using restriction enzymes EcoRV-HF (NEB) and MluHF (NEB).

2.1.2 Tissue culture techniques

2.1.2.1 Culturing cell lines

Media and culture conditions for cell lines used are summarised in Table 2.4. They were either cultured with Dulbecco's modified Eagle's medium (DMEM, Invitrogen), DMEM: Nutrient Mixture F-12 (DMEM/F12, Invitrogen), Rosewell Park Memorial Institute 1640 medium (RMPI, Invitrogen) or Ham's F12 Nutrient Mixture (Ham/F12) was supplemented with fetal bovine serum (FBS, Fetaclone III, Clonetech) and Penicillin Streptomycin (PS, Gibco) or Penicillin Streptomycin Glutamine (PSG, Gibco). Cells were maintained in a humidified incubator at 37 ºC and 5% CO_2 level throughout.

2.1.2.2 Cell transfections

MDA-MB-231 cells (ATCC) were transfected with a total of 5µg vectors using Lipofectamine (Life technologies) in Table 2.5. Transfected cells were maintained in a humidified incubator at 37ºC and at 5% CO_2 level for 48 hours. In the experiment expressing dCas9, Doxycycline (Clonetech) was used to induce the expression for 24 hours and cells were then selected by antibiotics.

Cell line	Medium	Supplements
MDA-MB-231	DMEM	10% FBS, 1% PSG
MDA-MB-157	DMEM	10% FBS, 1% PSG
HCC1569	RPMI	15% FBS, 1% PS
BT549	RPMI	10% FBS, 1% PSG
MCF10A	DMEM/F12	10% FBS, 1% PSG, 0.2% EGF, 0.5% Hydrocortisone, 0.1% Cholera Toxin, 1% Insulin
MCF7	DMEM	10% FBS, 1% PSG
T47D	RPMI	10% FBS, 1% PSG
SUM159T	Ham/F12	10% FBS, 1% PS, 0.1 Gentamicin, 5μg/ml Insulin, 1μg/ml Hydrocortisone
HS578T	DMEM	10% FBS, 1% PSG
HMLER	DMEM	10% FBS, 1% PSG

Table 2.4: Summary of the culture conditions for the breast cancer cell lines used

The medium and supplements used for the breast cancer cell lines in this project is shown in Table 2.4.

Experiment	Vectors	Lipofectamine	Antibiotics
gRNA and dCas9 targeting *BCL11A* regions	2.5µg gRNA-expression vector	Lipofectamine-3000	10µg/ml Blastocidin (Life Technology) for 7 days
	2.5µg Catalytically dead - expression vector		
	0.5µg Transposase		
Knock-out *BCL11A* targets	5µg px458 or px459	Lipofectamine LTX	1µg/ml Puromycin (InvivoGen) for 5 days
shRNA knock-down candidates	4.5µg shRNA vector	Lipofectamine LTX	400µg/ml Geneticin (Gibco) for 7 days
	0.5µg Transposase		

Table 2.5: Summary of cell transfections performed

The conditions for cell transfections in the three experiments mentioned above are shown in Table 2.5 in addition to the antibiotics and duration of selection used.

2.1.2.3 Proliferation assay

2000 cells of MDA-MB-231 transfected with shRNA plasmids were seeded. Cells were then washed with Phosphate-buffered saline (PBS) twice, trypsinised with Trypsin for 5 minutes at 37^{o}C and diluted with normal media to a total of 5ml. 10µl were pipetted into a cell counter and cells were counted every 2 days for 14 days. Each time-point was performed in triplicate.

2.1.2.4 Colony forming assay

MDA-MB-231 cells transfected shRNA plasmids were washed with PBS twice, typsinised with Typsin (Invitrogen) for 5 minutes at 37^{o}C and diluted with normal media to a total volume of 5ml. Cells were then filtered with 0.22µm filter (Millex) and 10µl were pipetted into a cell counter. 2000 cells were aliquoted into Eppendorf tubes, centrifuged at 400G for 5 minutes and supernatant removed. Next, 280µl of Matrigel (Corning) were added into the Eppendorf tubes on ice and cells were resuspended. 70µl of cell-Matrigel mixture were pipetted into a 6-well; this was performed in triplicates and 6-well plates were incubated at 37^{o}C, 5% CO_2 for 15 minutes to allow the Matrigel to solidify. 2ml of growth media were added and pictures were taken every 2 days from day 5. Images were overlapped with Photoshop and the number of colonies was counted with ImageJ.

2.1.3 Reverse transcriptase (RT)- polymerase chain reaction (PCR)

RNA was extracted with RNeasy Mini kit (Qiagen) following manufacturer's protocol and quantified with nanodrop. 1-2µg samples were made up to a 11µl mixture with 2 µl Random hexamer (Promega) and Nuclease-free water (Qiagen). Mixtures were then incubated at 65^{o}C for 5 minutes followed by 5 minutes on ice. 7µl of mastermix consisting of 0.5µl RT (Roche), 4µl 5xRT Buffer (Roche), 0.5µl RNasin (Promega) and 2µl 10mM dNTP (NEB) were added into the RNA mixture and cycled at 25^{o}C for 10 minutes followed by 42^{o}C for 10 minutes and lastly at 70^{o}C for 10 minutes. Finally, complementary DNA (cDNA) was diluted to 50µl with Nuclease-free water.

To perform quantitative polymerase chain reaction (qPCR), 2µl of cDNA were used per reaction. GoTaq® kit (Promega) and SYBR™ Select Master Mix (ThermoFisher) were used to measure gene expressions in samples. For reactions using the GoTaq® kit, mixtures were made up with 10µl of TaqMan solution, 1µl of probe for specific gene expressions, 2µl cDNA samples and 7µl Nuclease-free water in 96-well plates. Reactions with SYBR Green were made up with 10µl SYBR Green solution, 1.2µl primer pairs for specific gene expressions, 2µl cDNA samples and 6.8µl Nuclease-free water in 96-well plates. They were then placed into StepOnePlus Real-Time PCR machine and software (Applied Biosystems). Table 2.6 shows the genes of probes used for the qPCR experiments.

2.1.4 Protein extraction

Cells were grown in a confluent 6-well, washed with PBS twice and completely removed. 60µl of Radioimmunoprecipitation assay (RIPA) buffer (Cell Signalling Technology, CST) and 1x Protein inhibitor cocktail (Roche) were added and incubated on ice for 5 minutes. Cells were then scraped using a scraper (ThermoFisher), pipetted into an Eppendorf tube and a 27 gauge needle (Terumo) is used to shear the nuclei. Lysates were then incubated on ice for 30 minutes with brief vortexing every 10 minutes and centrifuged at 14,000G, 4°C for 10 minutes. Supernatant were transferred into a new Eppendorf tube and cell pellet discarded. Protein concentrations were determined with Pierce BCA kit (ThermoFisher) following manufacturer's protocol.

2.1.5 Western blot

Whole cell lysates were prepared with lysis buffer (RIPA) buffer and Protein inhibitors cocktail. Mixtures of 50µg protein samples, RIPA buffer and reducing loading buffer (X4 Laemmli buffer, Bio-Rad, 0.1M Dithiothreitol (DTT), Sigma) were heated at 95°C for 10 minutes and proteins were separated in acrylamide gels. Table 2.7 shows the composition of acrylamide gels used. Proteins in the gel were transferred onto Polyvinylidene difluoride (PVDF) membranes (Carl Roth) at 400mA for an hour. The membrane was washed with Tris-buffered saline (TBS)-Tween 20 (TBS-T) buffer (TBS buffer, 0.1% Tween-20, Sigma) once and blocked

Genes	Primers	Assay ID / Sequences	Manufacturer
Human *BCL11A*		Hs00256254_m1	ThermoFisher
Human *DEK*	Forward	TCCAAAGCCTTCTGGCAAACCATT	Sigma
	Complementary	TGGTGGCTCCTCTTCACTTTCTTTA	
Human *SSRP1*	Forward	CTCCTCGTGGTCGTTATGACA	Sigma
	Complementary	AGGGGATCTTGTAGTCAAAGGTC	
Human *PSIP1*	Forward	AAAACAGGGGTTACTTCAACCTC	Sigma
	Complementary	GGCCTTTCAGCATATTCCTTCT	
Human *GAPDH*		Hs02758991_g1	ThermoFisher
Human *PPP2R2D*	Forward	GGCAACGACTTCCAGTGGT	Sigma
	Complementary	ATGAGGGCGGCTTTTATTCTC	

Table 2.6: List of the expression of genes quantified in RT-qPCR

The Assay ID or primer sequences and manufacturer for genes quantified in RT-qPCR experiments are shown in Table 2.6.

Percentage of gels	Component	Volume
4% Stacking gel	30% Acrylamide (BioRad)	830 µl
	0.5M Tris-HCl, pH 6.8 (Merck)	630 µl
	10% SDS (To be filled in)	50 µl
	10% Ammonium per sulphate (Sigma)	50 µl
	TEMED (Sigma)	5 µl
	Deionised water	1ml
6% Resolving gel	30% Acrylamide	2 ml
	1.5M Tris-HCl, pH 8 (Merck)	2.5 ml
	10% SDS	100 µl
	10% Ammonium per sulphate	100 µl
	TEMED	8 µl
	Deionised water	5.3 ml
8% Resolving gel	30% Acrylamide	2.7 ml
	1.5M Tris-HCl, pH 8	2.5 ml
	10% SDS	100 µl
	10% Ammonium per sulphate	100 µl
	TEMED	6 µl
	Deionised water	4.6 ml

Table 2.7: The components for making Acrylamide gels in western blot experiments

The components and volumes used for the different percentage of Acrylamide gels in Western blots are summarised in Table 2.7.

in 5% non-fat milk powder (Marvel, The Premier Foods Group). Membranes were incubated with primary antibodies, shown in Table 2.8, overnight at 4ºC. Rabbit and mouse secondary antibodies (GE Healthcare) were probed and proteins were detected using Enhanced chemiluminescence (ECL) reagents (GR Healthcare).

2.1.5.1 Optimising Western blot conditions for antibodies – Protein DEK (DEK), Facilitates chromatin transactions (FACT) complex subunit SSRP1 (SSRP1) and PC4 and SFRS1 interacting protein 1 (PSIP1)

Optimisations have to be done for the primary antibodies for DEK, SSRP1 and PSIP1 prior to incubating with the western blots. The dilutants, 5% Milk or 5% Bovine serum albumin (BSA), used to dilute DEK, SSRP1 and PSIP1 primary antibodies were first optimised in HEK 293 cells, while keeping their concentration consistent. Figure 2.1 shows the resulting blots. DEK and PSIP1 primary antibodies work best when diluted with 5% Milk and SSRP1 primary antibodies with 5% BSA.

2.1.6 RIME

Cell culture of MDA-MB-231 was expanded to $240x10^6$ cells in complete DMEM growth media. Original media were removed, replaced with media containing 1% EM-grade formaldehyde (Polysciences) and crosslinked at room temperature for 8 minutes. Quenching was then performed using a final concentration of 1M Glycine solution. Cells were washed twice with ice-cold PBS (Life technology), harvested in PBS with Protease inhibitors cocktail and pellets collected by spinning down at 20,000 rcf for 5 minutes at 4ºC.

100µl of Protein G Magnetic beads (ThermoFisher) per immunoprecipitation were placed in 1.5ml Eppendorf tubes, washed with 1ml of 5mg/ml BSA in PBS solution and removed after being placed on a magnetic stand on ice for three times. Beads were resuspended in 500µl of the BSA/PBS solution with 15µg anti-V5 antibody (Abcam) or Rabbit IgG (Abcam) control antibody per immunoprecipitation. The tubes containing the beads and antibodies were rotated at 10 rpm in the cold room at 4ºC overnight.

Antibodies	Species	Concentration	Manufacturer
BCL11A	Rabbit	1:5000 in 5% Milk	Abcam
Anti-V5 tag	Rabbit	1:5000 in 5% Milk	Abcam
DEK	Rabbit	1:1000 in 5% Milk	Abcam
SSRP1	Mouse	1:100 in 5% BSA	Santa Cruz
PSIP1	Rabbit	1:250 in 5% Milk	Novus
Tubulin	Mouse	1:10,000 in 5% BSA	Abcam

Table 2.8: The antibodies used in western blot experiments

The species, concentration and manufacturer of the antibodies used in Western blots are summarised in Table 2.8.

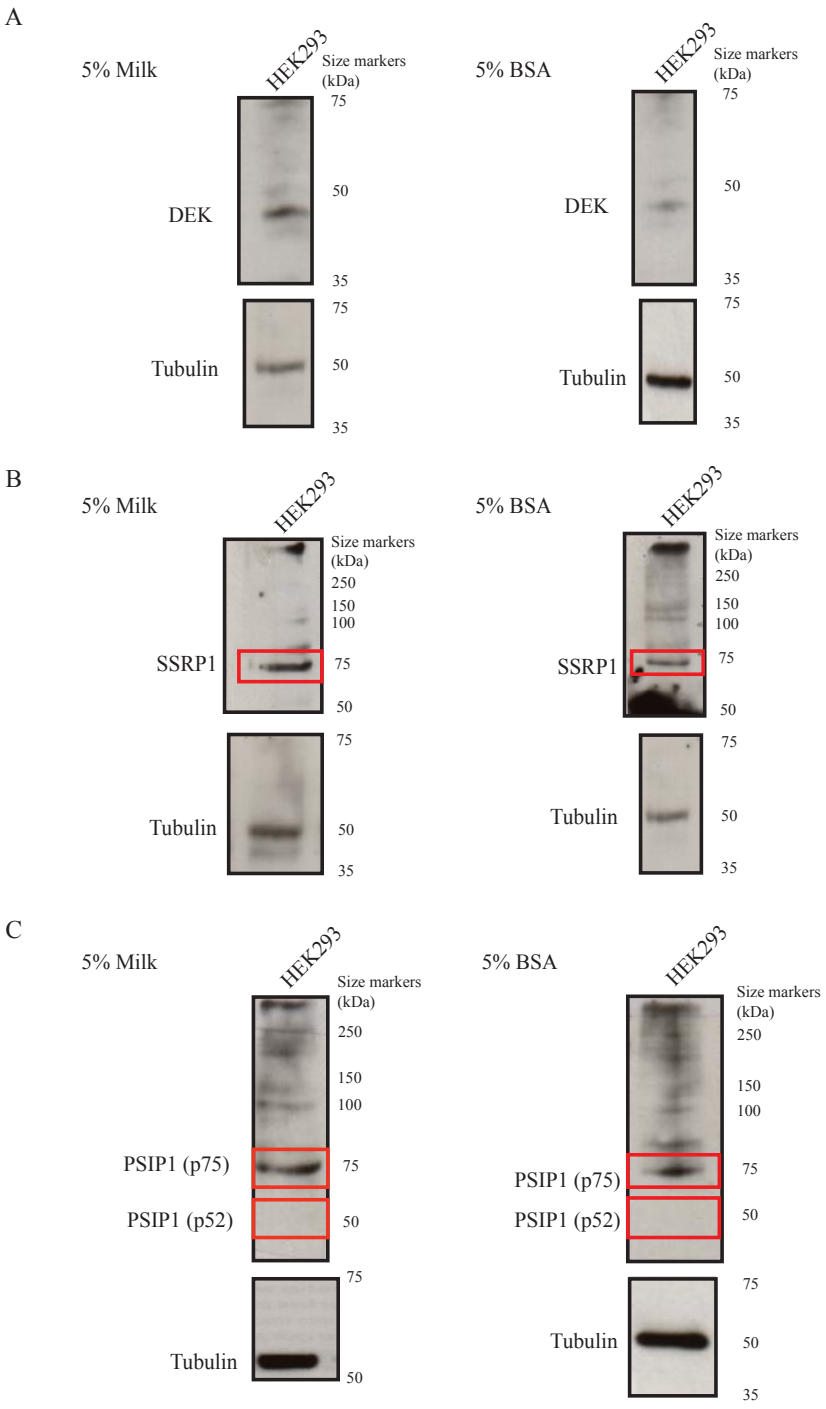


Figure 2.1: Optimisations for the conditions of western blot for primary antibodies, DEK, SSRP1 and PSIP1.

The condition to dilute DEK, SSRP1 and PSIP1 primary antibodies were optimised in HEK293 cells and Tubulin used as loading controls. DEK and PSIP1 blots showed best condition when diluted in 5% Milk and 5% BSA for SSRP1.

For each sample, nuclear fraction were extracted by resuspending cell pellets in 12ml LB 1 buffer (50 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)- potassium hydroxide (KOH) pH 7.5, Invitrogen, 140 mM Sodium Chloride (NaCl), Sigma, 1 mM EDTA, Sigma, 10% Glycerol, Sigma, 0.5% Tergitol-type NP-40 (NP-40), MP biomedical, and 0.25% Triton X-100, Promega) for 10 minutes at 4°C. Cells were pelleted and resuspended in 12ml LB 2 buffer (10 mM Tris- Hydrochloric acid (HCl) pH 8.0, Invitrogen, 200 mM NaCl, 1 mM EDTA, and 0.5 mM EGTA, Sigma) and incubated for 5 minutes at 4°C. Cells were pelleted again and resuspended in 3.6ml LB3 buffer (10mM Tris-HCl pH 8.0, 100mM NaCl, 1 mM, EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate, Sigma, and 0.5% N-lauroylsarcosine, Sigma, followed by water bath sonication at 4°C for 6 cycles of 30 seconds on, 30 seconds off. 30µl of 10% Triton X-100 was added to the samples and centrifuged for 10 minutes at 20,000 rcf to remove debris from lysate. The supernatant was incubated with antibodies-coated beads and rotated at 10 rpm in the cold room at 4°C overnight. On the following day, beads were washed ten times with RIPA buffer (50mM HEPES-KOH pH 7.6, 1mM Ethylenediaminetetraacetic acid (EDTA), 0.7% Sodium-deoxycholate, 1% NP-40, 0.5M Lithium Chloride, Sigma) and twice with 100mM ammonium hydrogen carbonate (AMBIC, Sigma). After the second AMBIC wash, beads were transferred into new Eppendorf tubes. Mass spectrometry analyses were performed at the proteomics facility at CRUK, CI.

2.1.6.1 Optimising sonication and determining reverse crosslinking conditions

2.1.6.1.1 MDA-MB-231 cells

Chromatin was sheared by sonication with a bioruptor pico at a rate of 30 seconds on, 30 seconds off per minute per cycle. Initial trial of sonication took 12 and 15 cycles with DNA purified with RNaseA (Roche) and Proteinase K (ThermoFisher) prior to reverse cross-linking with 5M NaCl at 65°C overnight, presented in Figure 2.2 A. The size of sonicated fragments was then determined on agarose gel. It was seen in Figure 2.2 A that the board smear which was far from the optimised sonicated fragments visualised as 200-400bp suggest uneven chromatin shearing and sonication was changed to a stronger bioruptor. 1 cycle from the new bioruptor

was equivalent to 6 cycles in previous optimisations. However, changing this step meant new optimisations of condition were required. The new sample in Figure 2.2 B was sonicated for 3 cycles and reverse-crosslinked for an hour at 65^{o}C. However, the unity of the size of chromatin sheared remains problematic for determining the optimised conditions. This was resolved by sonicating samples in Diagenode tubes rather than normal Eppendorf tubes. Figure 2.2 C tested 6 cycles of sonication and shearing was visualised in the dilutions of 1:8, 1:4 and 1:2. The uniform and optimal band sizes visualised in all of the dilutions suggest sonicating samples for 6 cycles is the optimal condition and was used for the real sonication.

2.1.6.1.2 HCC1569 cells

Optimisation of sonication was repeated with HCC1569 cells. Cells were collected and a range of sonication cycles were tested. DNA was then purified with RNaseA and Proteinase K before reverse crosslinking at 65^{o}C for an hour. Samples in Figure 2.2 D show the range of cycles tested. Among them, chromatin sheared with 7 cycles showed optimal fragment sizes of 200-400 base pairs and was taken forward for the real sonication.

2.1.7 Chromatin Immunoprecipitation

MDA-MB-231 or HCC1569 cells were expanded to 8.8x10^{6} cells in complete DMEM or RPMI growth media. Original media was removed and replaced with media containing 1% EM-grade formaldehyde (Polysciences) and crosslinked at room temperature for 10 minutes. Fixed cells were then quenched using a final concentration of 0.125M Glycine solution. Next, cells were washed twice with ice-cold PBS (Life technology), harvested in PBS with Protease inhibitors cocktail and pellets collected by spinning down at 20,000rcf for 5 minutes at 4^{o}C.

20µl of Protein G Magnetic beads per immunoprecipitation were placed in 1.5mL Eppendorf tubes, washed with 500µl of 5mg/ml BSA in PBS solution and removed after being placed in a magnetic stand on ice for three times. Beads were then resuspended in 250µl of BSA/PBS solution with 2.5µg Anti-V5 (Abcam); DEK

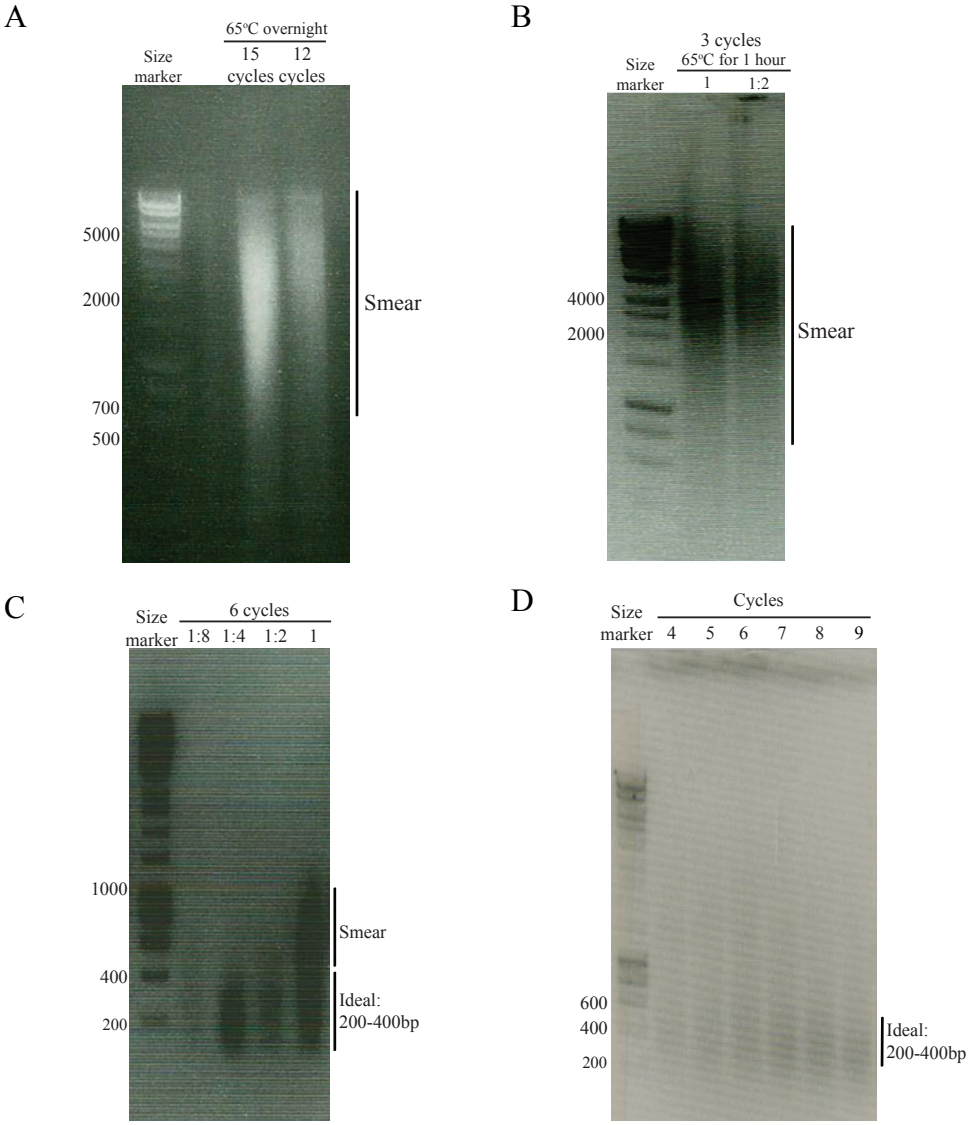


Figure 2.2: Optimising sonication and determining reverse crosslinking conditions for MDA-MB-231 and HCC1569 cells

Agarose gel in (A), 12 or 15 cycles and reverse-crosslinked at 65°C overnight. 15 cycles sonication produced more intense and longer smear, shifted from 2000-5000bp to 600-5000bp. (B) 3 cycles and reverse-crosslinked at 65°C for 1 hour. Smear of 2000-4000bp was observed. (C) 6 cycles sonication and reverse-crosslinked at 65°C for 1 hour. Uniform sheared fragments between 200-400bp was observed. (D) HCC1569 cells sonicated in 4-9 cycles and reverse-crosslinked at 65°C for 1 hour. 7 cycles sonication had the smear of the optimal fragmented sizes.

(Abcam); SSRP1 (Santa Cruz); PSIP1 (Novus); or 2.5µg controls of Rabbit IgG or Mouse IgG (Cell signalling) antibodies per immunoprecipitation. The tubes containing the beads and antibodies were rotated at 8 rpm in the cold room at 4^{o}C overnight. 10mM 0.5M EDTA, 1% SDS) or 25µl of input sample with 75µl elution buffer and incubated at 65^{o}C overnight, vortexing every 5 minutes for the first 15 minutes. On the following day, supernatant was transferred into new tubes and reverse crosslinked with first, 1µl of RNaseA at 37^{o}C for 30 minutes and second, 1µl Proteinase K at 55^{o}C for 2 hours. DNA was then eluted with MiniElute Kit (Qiagen) following manufacturer's protocol and quantified with qPCR.

For each sample, nuclear fractions were extracted by resuspending cell pellets in 1ml LB 1 buffer for 10 minutes at 4^{o}C. Cells were pelleted and resuspended in 1ml LB 2 buffer and incubated for 5 minutes at 4^{o}C. Cells were pelleted again and resuspended in 300µl LB 3 buffer followed by water bath sonication at 4^{o}C for 7 cycles of 30 seconds on, 30 seconds off. 30µl of 10% Triton X-100 was added to the samples and centrifuged for 10 minutes at 20,000 rcf to remove debris from lysate. The supernatant made 700µl RIPA (50mM Hepes KOH, Invitrogen, 500mM 8M LiCl, 1mM 10% NP-40, MP biomedical, and 0.7% Sodium-deoxycholate, Sigma), once with 700µl Tris-EDTA (TE)/NaCl buffer and spin to remove residual buffer. Beads were resuspended in 100µl elution buffer (50mM 1M Tris HCl, 10mM 0.5M EDTA, 1% SDS) or 25µl of input sample with 75µl elution buffer and incubated at 65^{o}C overnight, vortexing every 5 minutes for the first 15 minutes. On the following day, supernatant was transferred into new tubes and reverse crosslinked with 1µl of RNaseA at 37^{o}C for 30 minutes and next 1µl Proteinase K at 55^{o}C for 2 hours. DNA was then eluted with MiniElute Kit (Qiagen) following manufacturer's protocol and quantified with qPCR.

To perform qPCR, 2µl of samples were used per reaction. SYBR^{TM} Select Master Mix (ThermoFisher) was used to test the specificity of ChIP. Reaction mixtures were made up of 10µl of SYBR Green, 1.2µl of primer pair, 2µl cDNA samples and 6.8µl Nuclease-free water in 96-well plates. 96-well plates were placed into StepOnePlus Real-Time PCR machine and software (Applied Biosystems). Table 2.9 shows primers used to test the specificity of ChIP experiments.

Primers	Sequences
Primer 1 forward	ACACCCAGGAGCCAGCAG
Primer 1 complementary	TGTGTGCATTAAAGAGACAGGG
Primer 2 forward	AAGCAGACCAGCTCAAGTGAA
Primer 2 complementary	GCGTACACAGCTTTCTTTCCT
Primer 3 forward	TGTTAAACCAGCCTTGGAGTCA
Primer 3 complementary	TCACTATTGTTACCTGAAGGGG
Primer 4 forward	GGAAATCCAGACCTTGGCAGT
Primer 4 complementary	GGGTGAGCCTGCGGTTATTG

Table 2.9: The primer sequences used in ChIP-qPCR experiments

The sequences of the primers used in ChIP-qPCR experiments are shown in Table 2.9.

CHAPTER 3

Designing a novel system to target *BCL11A*'s genomic regions with the CRISPR-Cas9 system

3.1 Introduction

The breast cancer subtype triple-negativebreast cancer (TNBC) is the focus of this study. It shows clinical characteristics of a higher mortality rate and more frequent recurrence within five years than are found with other subtypes [42]. In addition, this subtype does not express hormonal receptors for Oestrogen, Progesterone or the Human epidermal growth factor receptor 2 (HER2). The lack of these receptors in TNBC patients leads to their ineligibility for treatments that have been developed to target these receptors, leading to a poor clinical outcome. Therefore, there is an urgent need to find new molecular targets so that these patients can be treated.

One of the advances in the study of this subtype of breast cancer was the discovery of the B-cell lymphoma/leukaemia 11A (*BCL11A*) gene as an oncogene [74]. The *BCL11A* gene was found to show high expression in TNBC patients. A previous study by our group has shown its importance in the initiation and maintenance of tumourgenesis [74]. However, the regulation of this oncogene in TNBC remains unknown. On the other hand, the functional roles of *BCL11A* in other biological systems are better understood. It is known that *BCL11A* is required: during foetal development and for cell functions, including γ- to β-globin gene switching [62] [63] [65]; to regulate axon outgrowth and branching during brain development [71][119]; and in the formation of immune cells such as B-cells [66]. Due to the importance of *BCL11A* in these roles, an inhibitor which globally targets *BCL11A* is not desirable, as it would affect normal cell function and even cell survival. One way to overcome this problem is to identify specific regulatory pathways of *BCL11A* in TNBC and to develop compounds that target the interactions between these regulatory gene networks or protein complexes.

To investigate this regulation of *BCL11A* in TNBC, we have employed the clustered regularly interspaced palindromic repeats (CRISPR) system, which was originally discovered as a defence mechanism for bacteria and archea [89] [90]. This project implements a novel design in dCas9 is used instead of the normal Cas9. This

enables specific sites of *BCL11A* to be targeted while avoiding DNA double-strand breaks that would otherwise be introduced by normal Cas9.

The second novelty in this design is the use of dCas9 as a protein to be pulled down by specific antibodies. This has enabled the study of transcriptional regulators and the identification of those that may have roles in the regulation of *BCL11A*'s expression at our sites of interest.

Therefore, the studies detailed in this chapter aim to design and evaluate the effectiveness of the CRISPR-dCas9 system to target *BCL11A*'s genomic regions.

3.2 Results

3.2.1 The determination of which *BCL11A* genomic regions to study and the design of guide RNAs for expression plasmids

To employ the CRISPR-Cas9 system to target specific sites of *BCL11A*, the design of gRNA was first considered. These gRNAs are responsible for the guidance of dCas9 to the *BCL11A* sites of interest. The sequences of the human *BCL11A* messenger RNA (mRNA) transcript 1 were obtained from the University of California Santa Cruz (UCSC) genome browser (Accession number: NM-022893) [117]. The UCSC genome browser is an open-access tool available to search for annotated genomic sequence data [120]. Through the use of this tool, the sites of *BCL11A* with active transcription based on the annotations represented by the H3K27Ac histone marker and sensitivity to DNaseI were identified. The H3K27Ac histone marker was used to distinguish between active and inactive gene regions [121], while sensitivity to DNaseI indicated regions to which transcription factors could bind [122]. These features pinpointed sites with potential active transcription. The positions of these genetic sites of interest at the 5' and 3' regions are shown in Figure 3.1, and they are known in this study as Site A and Site B. The sequences of these sites were then used to design gRNAs and their complementary sequences using the CRISPR-Cas9 website: http://crispr.mit.edu/ [102].

3.2.2 The importance of *BCL11A*'s genomic targets for cell survival

To investigate whether the appropriate genomic regions of *BCL11A* had been targeted, plasmids that expressed both gRNA and Cas9 were co-transfected into MDA-MB-231 cells. The main reason to perform experiments in this cell line was its establishment back in the 1970s [123] and as a standardised cell line for researchers to study TNBC.

The Cas9 was used to delete small fragments of Sites A and B. The positions of these gRNAs are shown in Figure 3.2 A. The two plasmids used are shown in Figure 3.2 B, both of which express Cas9. The plan was that they would each target one

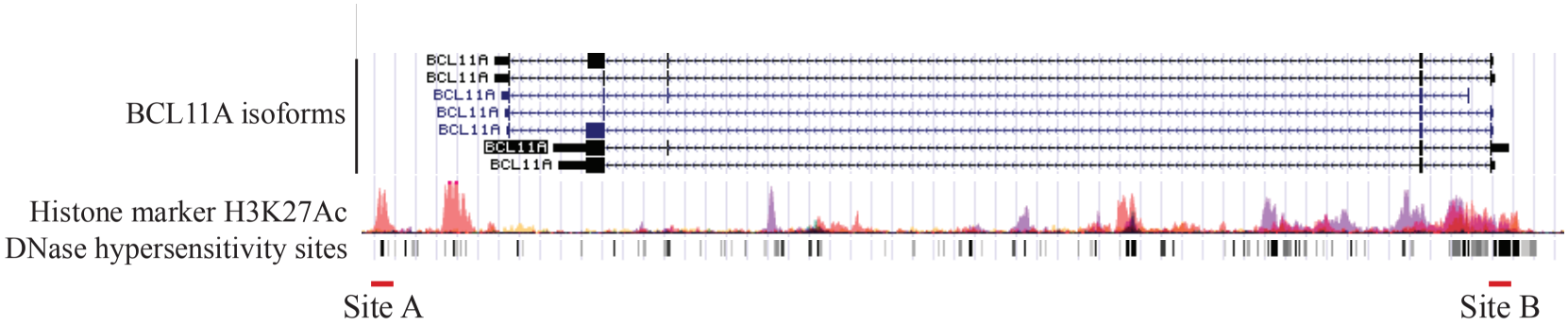


Figure 3.1: Designing gRNA complementary to designated sites.

Human *BCL11A* mRNA transcript 1 was used to design gRNA at specific sites displaying active transcription [117]. These are represented by the histone marker H3K27Ac and sensitivity against DNase I. The red lines are graphic representations of the positions of the gRNA aon the transcript and these are named Sites A and B.

gRNA site and guides the Cas9 to these targets, resulting in genome deletion. For example, one plasmid might target the gRNA 1 site of Site A while the other might target the gRNA 3 site of Site A. Co-transfection of these two plasmids would result in the deletion of regions between sites 1 and 3.

In addition to expressing gRNAs and Cas9, the plasmids also expressed different selection markers; one expressed enhanced green fluorescent protein (EGFP) signals while the other expressed a puromycin-resistance gene. These markers enabled the selection of cells co-transfected with both plasmids, firstly with puromycin and secondly by fluorescence-activated cell sorting (FACS) for single cells with EGFP expressions. The experimental plan was to identify homozygous clones and collect cells to examine their effects on the expression of *BCL11A.* The schematic of this plan is shown in Figure 3.2 C.

However, transfected cells did not survive the selection by puromycin. The two-plasmid Cas9 deletion system had been used previously by our group, so the chance that cell death was caused by the introduction of plasmids into the cells could be ruled out. Possible explanations for the result were that these genomic regions were important for the expression of *BCL11A,* or they affected the downstream pathway, and therefore they were detrimental to cell survival. This proved that Sites A and B were important and the experiment could proceed.

3.2.3 The generation of stable cell lines to express gRNAs and dCas9

After the determination of the importance of the regions of *BCL11A* to be studied, gRNAs whose designs were discussed in section 3.2.1 were cloned into the gRNA-expression plasmid and transfected alongside the dCas9-expression plasmid into MDA-MB-231 cells. The key features of these plasmids are represented in Figures 3.2 A and B.

Each gRNA designed with the CRISPR-Cas9 website consisted of short sequences of 20 nucleotides followed by a protospacer adjacent motif (PAM) sequence, which took the form NGG (a nucleobase followed by two guanine nucleobases). NGG

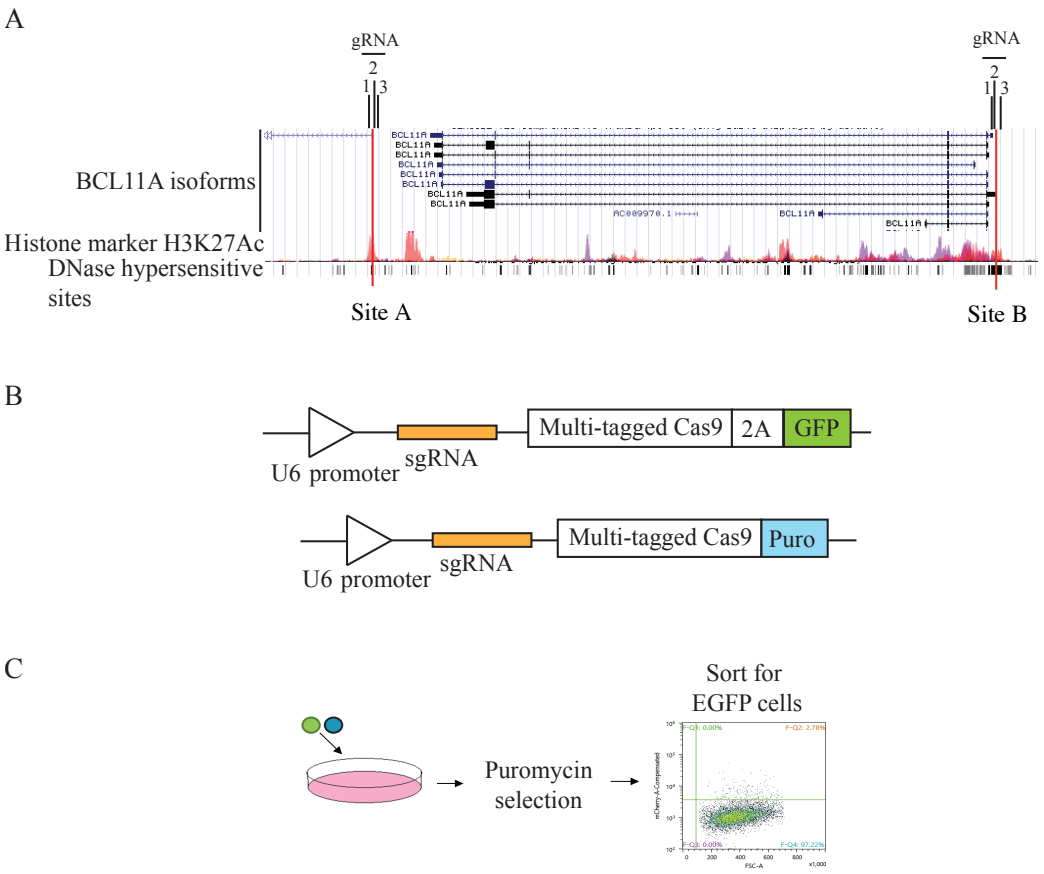


Figure 3.2: Deleting Sites A and B with Cas9 expression plasmids.

(A) shows positions of primer pairs designed to delete small neighbouring regions of *BCL11A*'s target sites. These sites displayed active transcription shown by the histone marker H3K27Ac and hypersensitivity to DNaseI. The red lines are graphic representations of target sites – Sites A and B. (B) shows the linearised form of plasmids expressing both gRNA and Cas9. gRNA expression was driven by the human U6 promoter. In addition, Cas9 was co-expressed either with an EGFP fluorescence signal through a self-cleavable 2A protein or with a Puromycin resistance gene. (C) shows the schematic of the experimental plan. Cells were first transfected with the Cas9 plasmids, followed by puromycin selection and then single sort for cells expressing EGFP. Cells will then be collected and the effect on *BCL11A*'s expression will be investigated.

sequences specific for humans [94] were annealed to form primers complementary to the sites of interest. These were then cloned into the cloning vector shown in Figure 3.3 A. The cloning vectors were supplied by Dr Xuefei Gao of the Sanger Institute, Cambridge. Expression was driven by the human U6 promoter. Cells were then co-transfected with the dCas9-expression plasmid in Figure 3.3 B.

The genome transfer was achieved using the PiggyBac (PB) transposon system shown in Figure 3.3 C. The enzyme transposase was employed to recognise regions of the plasmids that contained inverse terminal repeats (ITR), and to cut and paste these into a genomic region with TTAA sequence [124]. This method offered simple spatial control. Through the use of this PB gene delivery system, gRNAs were expressed with mCherry fluorescence and the blasticidin-resistance gene. These features enabled cells that expressed this plasmid to be selected with blastocidin followed by FACS. A summary of this experimental approach is shown through the use of a schematic in Figure 3.4 A. However, a problem emerged at the FACS stage; the percentage of cells that showed double-positive fluorescence signals was found to be very low. Therefore, optimisations of this experimental approach were needed.

The schematic for the new experimental plan is shown in Figure 3.4 B. Through use of a similar method to that shown in Figure 3.4 A, cells were first transfected with gRNA-expression plasmid followed by blasticidin selection and FACS for cells that expressed gRNA. These were represented by the mCherry fluorescence signal. Next, these cells were transfected with the dCas9-expression plasmid. As was observed in the case of the gRNA-expression plasmid, this dCas9 plasmid also showed an EGFP fluorescence signal, which enabled the sorting of cells through FACS and the generation of stable cell lines that expressed both gRNA and dCas9 plasmids. The percentages of cells that expressed both plasmids, which were produced using the two strategies, were compared. Figures 3.4 C and D are graphs comparing the numbers of cells that were double-positive and that expressed gRNA for Site A and Site B, respectively. The adoption of this two-step transfection system improved the production rate of cells co-transfected with both plasmids from 0.04 per cent to eight per cent and six per cent, respectively.

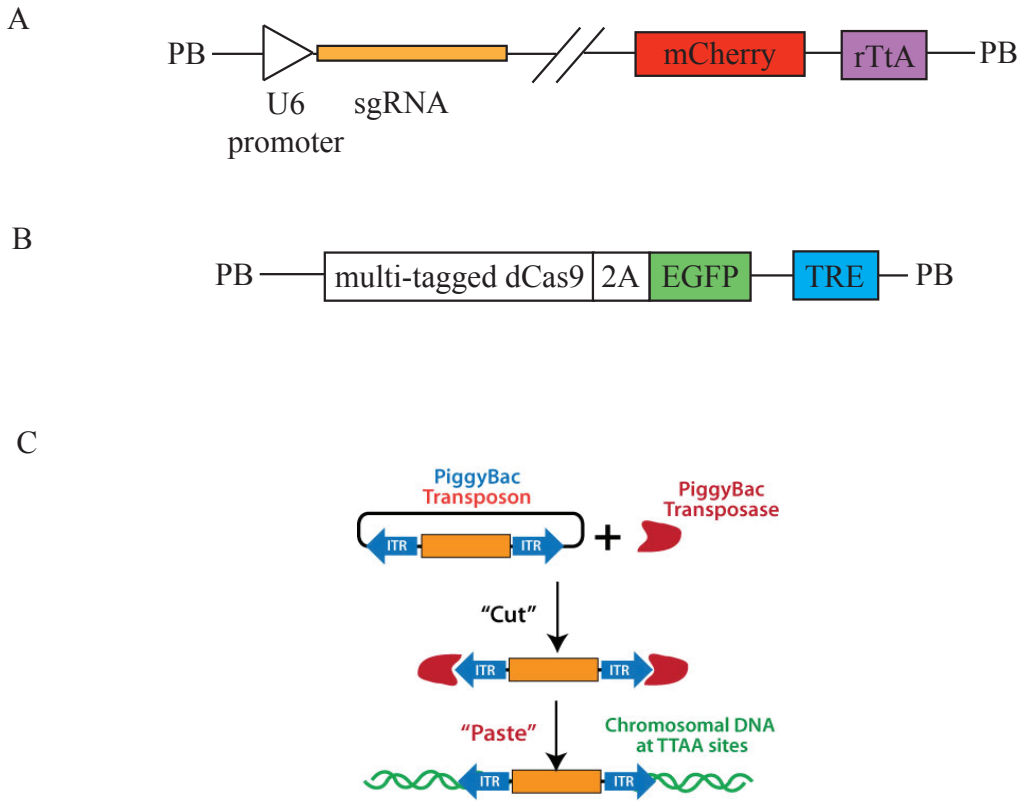


Figure 3.3: gRNA- and dCas9-expressing vectors and gene delivery.

Graphic representations in (A) and (B) show the linearised forms of gRNA- and dCas9- expressing plasmids. In the gRNA-expressing plasmid, gRNA expression was driven by the U6 promoter. It also expresses the mCherry fluorescence signal and rTtA component of the Tet-On system. Meanwhile, multi-tagged dCas9 was co-expressed with EGFP signal through a self-cleavable 2A protein in between. It also has the TRE component of the Tet-On system to initiate transcription. Both plasmids employ the PB transposon and Tet-On system. (C) PiggyBac transposon transfer specific vector regions using a cut-and-paste mechanism. The enzyme, transposase, recognises and cuts the the inverted terminal repeats (ITR). This enables an easy cut-and-paste mechanism for transferring the region of gRNA inserted into a specific TTAA site. (Modified from [125])

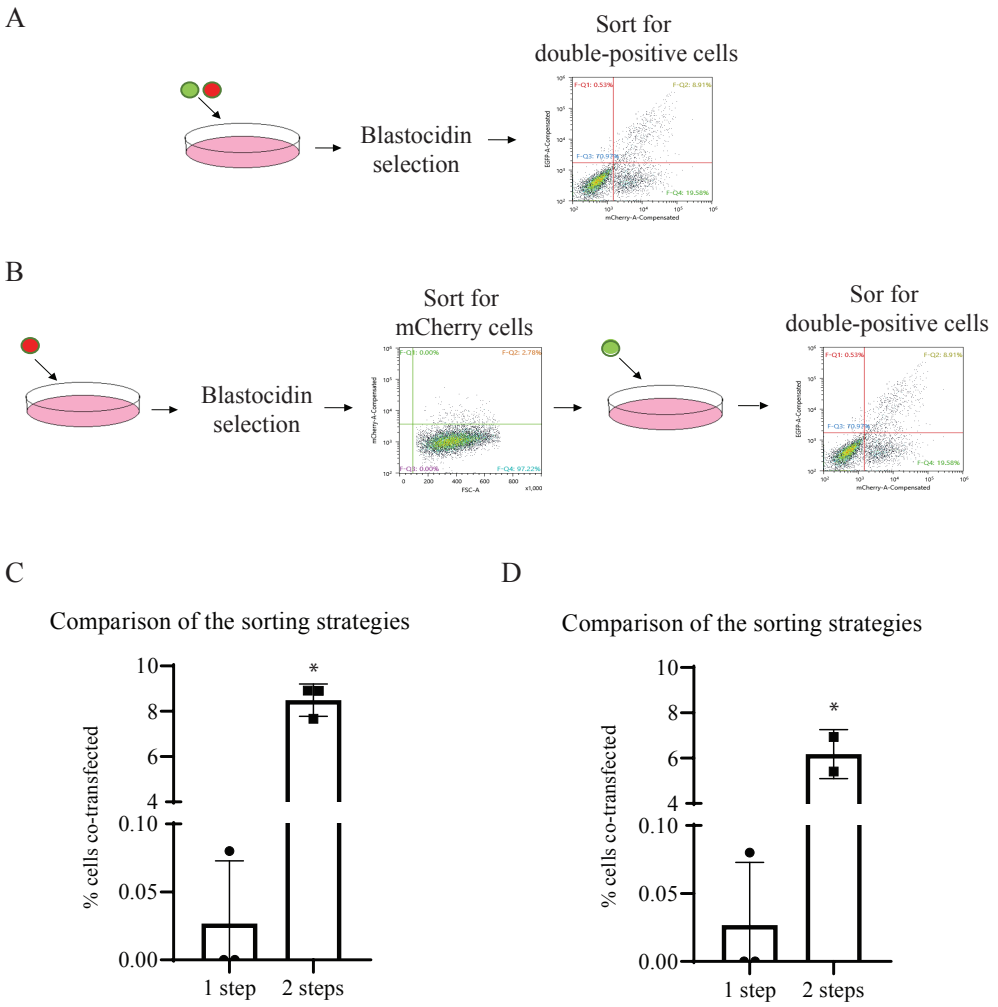


Figure 3.4: Optimisations of co-transfections.

The schematics of transfections are shown in (A) and (B). (A) shows the 1-step experimental plan where MDA-MB-231 cells were first co-transfected with both plasmids followed by Blastocidin selection and sort for cells which are double-positives. (B) shows the new schematic of the experimental plan. Cells are now transfected with firstly gRNA-expression plasmids, followed by Blastocidin selection and sort for cells with mCherry. Next, these cells are transfected with the dCas9-expression plasmid and sorted for double-positives. (C&D) describe the comparison of cells with double-positives from the two schemes. (C) are cells with gRNA for site A and (D) is for site B. (* <0.05, t-test, n=2-3.)

3.2.4 The validation of the induction and expression of dCas9 using the Tet-On system

The transposon system was used for stable gene delivery, while the Tet-On system was employed to induce dCas9 expression with Doxycycline. This was achieved by including a component of the Tet-On system [126], the recombinant tetracycline-controlled transcription factor (rTtA), in the gRNA-expression plasmid in addition to gRNA, the mCherry fluorescence signal and the blasticidin-resistance gene. This rTtA is normally inactive but constitutively expressed in the cells. Activation only occurs when Tetracycline or its derivative, Doxycycline, is introduced into the culture medium. The activated rTtA protein then binds to the Tet response element (TRE) in the dCas9-expression plasmid, and this enables temporal control of the expression of dCas9.

3.2.4.1 Determination that dCas9 is inducible and its impact on *BCL11A* expression investigated

The fluorescence signals in the expression plasmids enabled the expression of gRNA and dCas9 to be validated. The results are shown in Figure 3.5 A. The EGFP from the dCas9-expression plasmid was expressed at the same time as the protein and then separated shortly afterwards through the use of a self-cleavable 2A protein. To assess whether this co-expression system was functional, RNA and proteins were collected to perform a qPCR and western blot analysis, respectively. The western blot shown in Figure 3.5 B indicated that cells had been co-transfected with the gRNA- and dCas9-expression plasmids. Anti-V5 tag antibody was used against the V5 from the multi-tagged dCas9, and tubulin was employed as a loading control. The level of dCas9 observed in the non-induced samples was measured to provide a background level. Inductions were clearly seen in the samples that contained Sites A and B. Thus, as expected, inducing the expression of dCas9 protein with Doxycycline correlated with the observed levels of EGFP.

After validating the expression of dCas9, it was necessary to determine whether this induction affected the expression of *BCL11A.* Data obtained from qPCR analysis,

shown in Figure 3.5 C, suggested that dCas9 expression did not show any impact on the expression of *BCL11A* at either Site A or Site B. The cells described above were collected after 24 hours of Doxycycline induction. To determine whether increased time of exposure to Doxycycline would affect expression of *BCL11A,* the duration of the dCas9 induction experiment was extended to 48 hours. The result is illustrated in Figure 3.5 D and shows that the expression of dCas9 did not affect the expression of *BCL11A* regardless of the duration of induction. This indicates Cas9 bind close to Sites A and B but does not obstruct access of transcription factors to these sites that regulates *BCL11A*'s expression.

3.2.5 Determination that dCas9 binds specifically at its corresponding targets

In addition to validating the expressions of the plasmids, it was important to validate the specificity of the binding of dCas9 using ChIP coupled with qPCR. In the ChIP experiment, samples that were not subjected to Doxycycline induction were used to provide background data, while samples pulled down with IgG antibody instead of dCas9 were employed as a control. In the qPCR step, two primers specific for Sites A and B were also used. These were named primer 1 and primer 2 and their positions on *BCL11A* are shown in Figures 3.6 and 3.7. Using these primers, dCas9 was found to be bound at both target sites shown in Figures 3.6 B and 3.7 B. The specificity of dCas9 was also confirmed through the use of opposite primers as non-specific controls: that is, primer 2 was used for samples with gRNA for Site A and primer 1 for Site B.

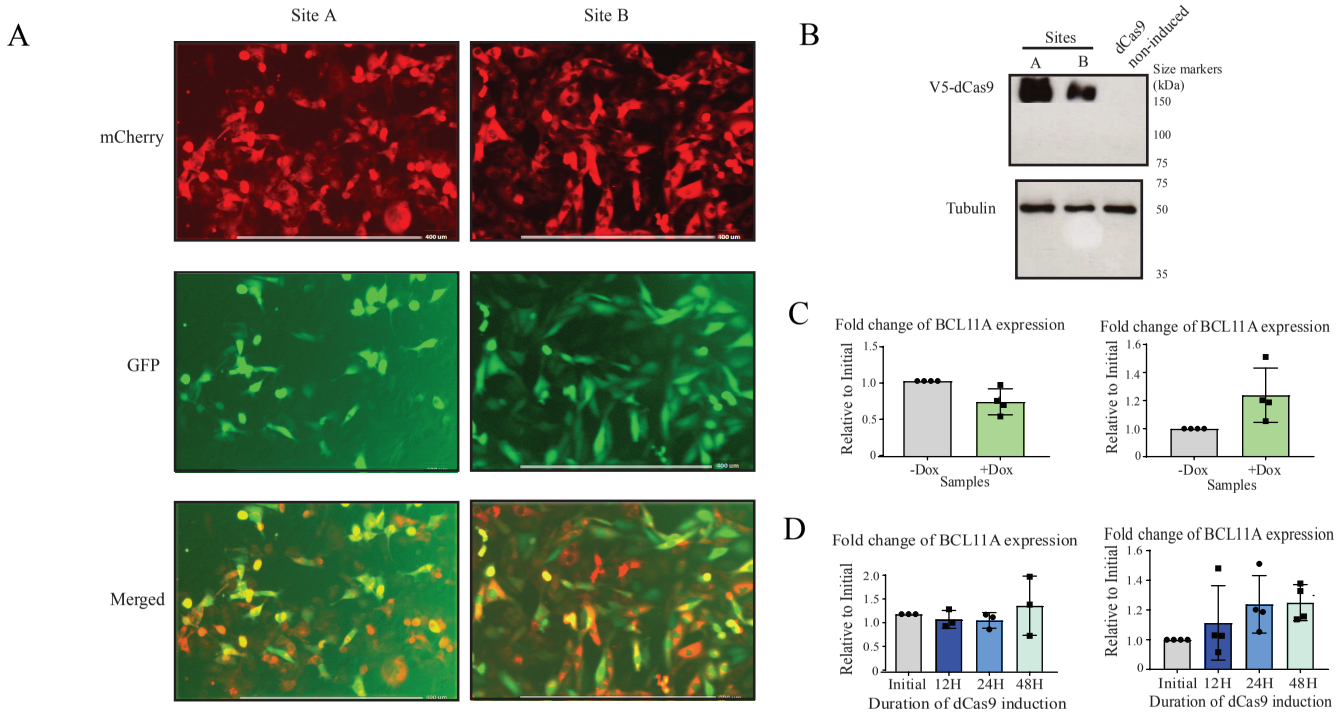


Figure .3.5: Effect on *BCL11A* expression with dCas9 induction for 24 and 48 hours.

(A) shows the panel of mCherry and EGFP fluorescences in cells with 5' or 3' gRNA and dCas9 plasmids at 10x magnification. The two fluorescence signals represent the level of gRNA- and dCas9- expressed in these cells. Overlap is indicated by yellow signals in the merged panel. (B) shows the Western blot of dCas9 expression and Tubulin as a loading control. RT-qCPR data demonstrates the effect on *BCL11A* expression after inducing dCas9 expression for 24 (C) and 48 (D) hours. n= 3-4

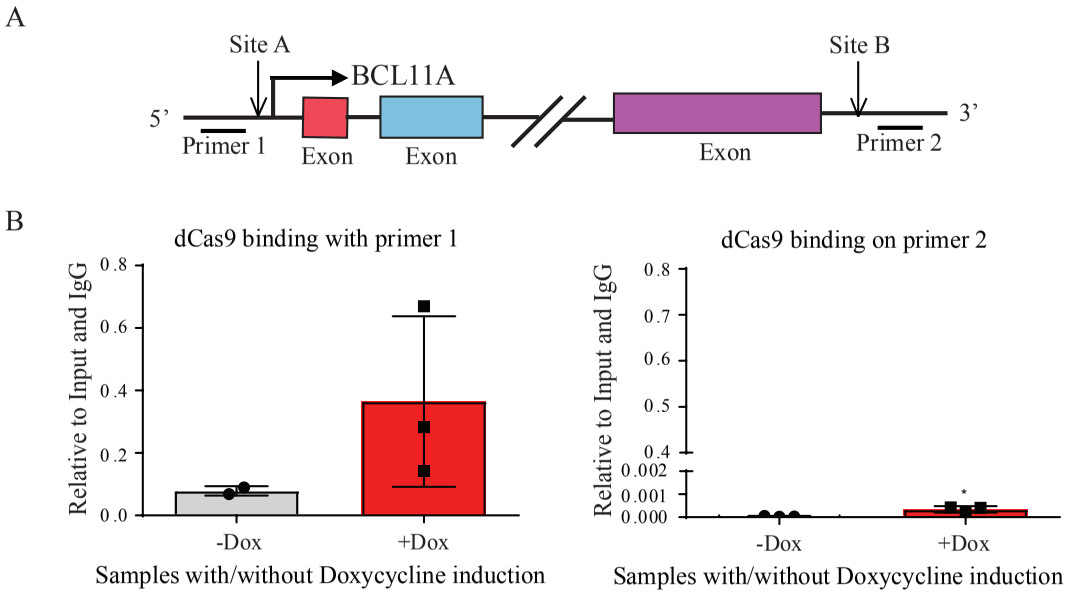


Figure 3.6: ChIP on samples with gRNA for site A showed specific dCas9 binding.

(A) is a graphic representation of the positions of Sites A and B and the primers used in the qPCR experiment. Following dCas9 pull-down, primer 1 and 2 designed to be as close to Sites A and B as possible were used in the qPCR step. Anti-V5 tag against the multi-tagged dCas9 was used in (B). Samples without Doxycycline induction were used as internal controls and graphs are plotted in relation to input and IgG controls. (*<0.05, t-test, n=3)

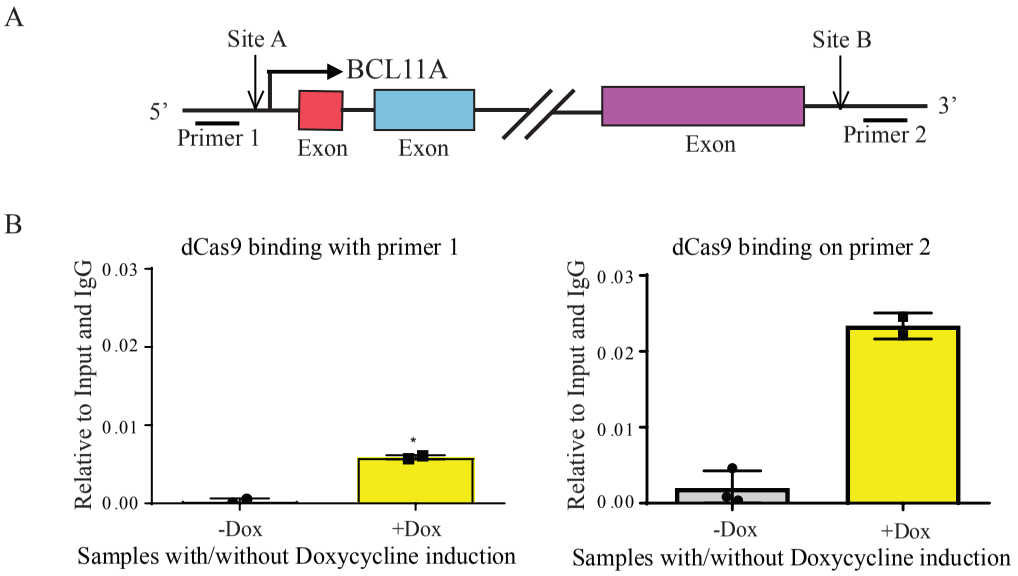


Figure 3.7: ChIP on samples with gRNA for site B showed specific dCas9 binding.

(A) is a graphic representation of the positions of Sites A and B and the primers used in the qPCR experiment. Following dCas9 pull-down, primers 1 and 2 designed to be as close to Sites A and B as possible were used in the qPCR step. Anti-V5 tag against the multi-tagged dCas9 was used in (B). Samples without Doxycycline induction were used as internal controls and graphs are plotted in relation to input and IgG controls. (*<0.05, t-test, n=2 (Average of two sets of data))

3.3 Discussion

This chapter has described the design of gRNA for the CRISPR-Cas9 system and the experimental method used to target particular genetic regions. Specific genomic regions on the *BCL11A* gene that displayed active transcriptional regulation were selected for the gRNA design. These gRNAs were then cloned into gRNA-expression plasmids, which guided dCas9 to bind to their corresponding targets. Plasmids were co-transfected into MDA-MB-231 cell lines; the expression of dCas9 and its specificity regarding its ability to bind to target sites was then validated.

3.3.1 The evaluation of the importance of chosen target sites

The sites of interest on *BCL11A*, which have been designated as Sites A and B, were chosen because of their strong regulation of transcription as displayed in the UCSC genome browser [117]. This was represented by their activity towards the H3K27Ac histone marker [121] and their sensitivity to DNaseI [122]. To investigate if these target sites were indeed important for expression of *BCL11A,* Cas9 was used to delete Sites A and B. However, the cells did not survive the initial deletion of the whole genetic regions. To overcome this problem, deletion of smaller fragments within the regions was employed, to investigate whether any regions within Sites A and B played a major role in the regulation of the expression of *BCL11A*. However, the cells also did not survive this experiment after puromycin selection. This led to the conclusion that these target regions might be important for gene expression, or the deletions affected the downstream pathway and thus were detrimental to cell survival.

3.3.2 The validation of components of the experimental design

After establishing the importance of the regions to be studied, cloning was resumed of the designed gRNA into the gRNA-expression plasmid. The altered plan to transfect separately the gRNA- and dCas9-expression plasmids significantly increased the percentage of cells that expressed both plasmids. Once the components of the experimental design had been validated, such as dCas9

expression and its binding specificity, these cell lines were expanded in cell cultures and fixed by formaldehyde for the RIME experiment described in Chapter 4.

I attempted to expand this work beyond the identification of transcriptional regulators for *BCL11A* in TNBC to generate stable cell lines for other breast-cancer subtypes such as MCF7 and T47D. However, the cells did not survive blasticidin selection or were not transfected. This was demonstrated by the very small number of cells collected during the FACS stage.

A key component of the experimental design that required validation was the PB transposon system. The mCherry and EGFP fluorescence signals showed in Figure 3.5 A suggested that gene delivery by this system was successful. The protein level of dCas9 expression was also examined by the use of western blots. The western blot displayed in Figure 3.5 B indicated successful induction of dCas9 by the introduction of Doxycycline into the culture medium used in the Tet-On system.
As well as the validation of the expression of gRNA and dCas9, it was also important to investigate whether the induction of dCas9 had any impact on expression of *BCL11A*. This was performed through RT-qPCR experiments. The initial induction experiment was performed for 24 hours, during which protein lysate were collected at various time-points. No change in the expression of *BCL11A* was observed after induction of dCas9 expression at both Sites A and B. This effect was also seen when the duration of induction was extended to 48 hours.

The final evaluation required of this CRISPR-Cas9 system was of the specificity of dCas9. V5-tagged antibody was used to pull-down multi-tagged dCas9 during ChIP experiments. The site of dCas9 binding was then investigated through the use of primers specific to Sites A and B during the qPCR step. The data shown in Figures 3.6 and 3.7 indicated that dCas9 binding was specific upon induction.

In summary, the experimental design was evaluated and cell lines were generated which expressed gRNA for Sites A and B and dCas9 in a stable manner. These cells were then prepared for the RIME experiment described in Chapter 4.

CHAPTER 4

Proteomic analysis

4.1 Introduction

RIME is a technique used to study proteomics. It combines ChIP with mass spectrometry. The coupling of these two techniques enables the use of RIME to study and identify proteins positioned at genomic sites that might regulate *BCL11A* in Triple-negative breast cancer. This process was applied after cells were transfected with the adapted CRISPR-Cas9 system as described in Chapter 3. It is presented in Figure 4.1.

The aim of this chapter is to analyse the RIME pull-down and to validate the use of this technique by determining whether these proteins are biologically relevant and show important functions in the regulation of *BCL11A*.

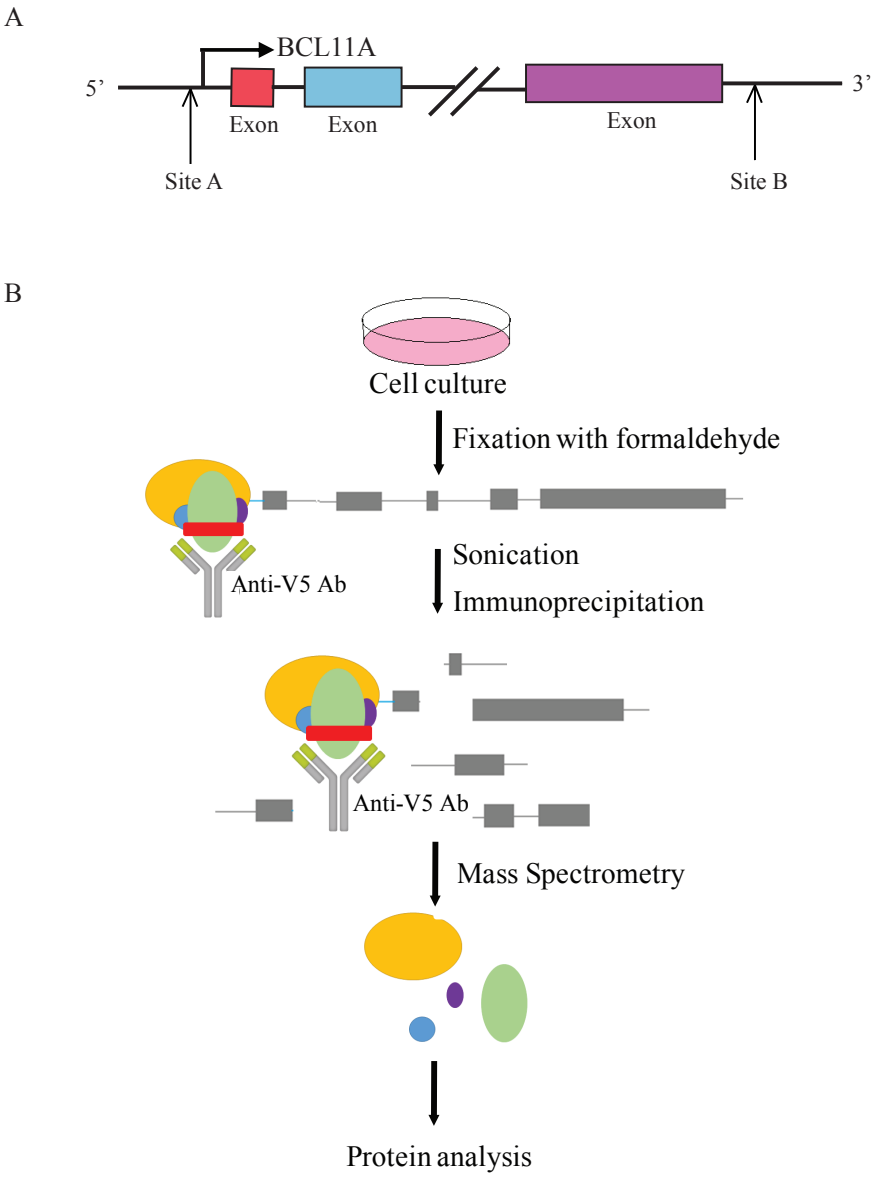


Figure 4.1: Experimental work flow for RIME.

The graphic representation in (A) shows the positions of our target sites, Sites A and B, on *BCL11A*. (B) shows the schematic of the experimental work flow for RIME. Cells were first transfected with the CRISPR-Cas9 system, described in Chapter 3, and then fixed with formaldehyde and sonicated to shear DNA. Next, samples were processed as a ChIP experiment using Anti-V5 antibodies to pull-down Cas9 itself along with proteins bound in close proximity. These protein partners were then identified with mass spectrometry and protein analysis.

4.2 Results

4.2.1 Formation of a reverse crosslink for RIME to visualise sizes of sonicated chromatin

The establishment of stable cell lines was described in Chapter 3. The number of lines was expanded to twleve 15cm plates. These were fixed with formaldehyde to crosslink proteins at close proximity and to stabilise interactions [83]. This is an important step in order to study protein-protein interactions in designated genetic regions. To investigate proteins with the potential to regulate the expression of *BCL11A,* nuclear fractions from the cell pellets were isolated. Chromatin was then sheared by sonication. Chapter 2: Materials and Methods provides details of the processes and optimisations employed.

4.2.2 The analysis of protein pull-down from the RIME dataset

The number of proteins collected during the initial pull-down in RIME was 326 from Site A and 491 from Site B. These figures were narrowed down using multiple parameters to reduce the number of proteins required to be studied, refers to Appendix A for the full protein list after each filter. The work flow is shown in Figure 4.2.

Firstly, the RIME experiment was performed in triplicate for each sample and the pulled-down proteins were mixed together. The overlapping members of the list of proteins pulled down using Anti-V5 antibodies against the multi-tagged dCas9 were compared against those pulled down in the control IgG sample, in which the Anti-V5 antibody had been replaced with an IgG antibody. Tables 4.1- 4.3 list the top 20 proteins that remained after this elimination between or from Sites A and B.

Secondly, the list of the remaining proteins was compared with those observed between and from the two targets. Proteins in common are shown on the Venn diagram in Figure 4.2.

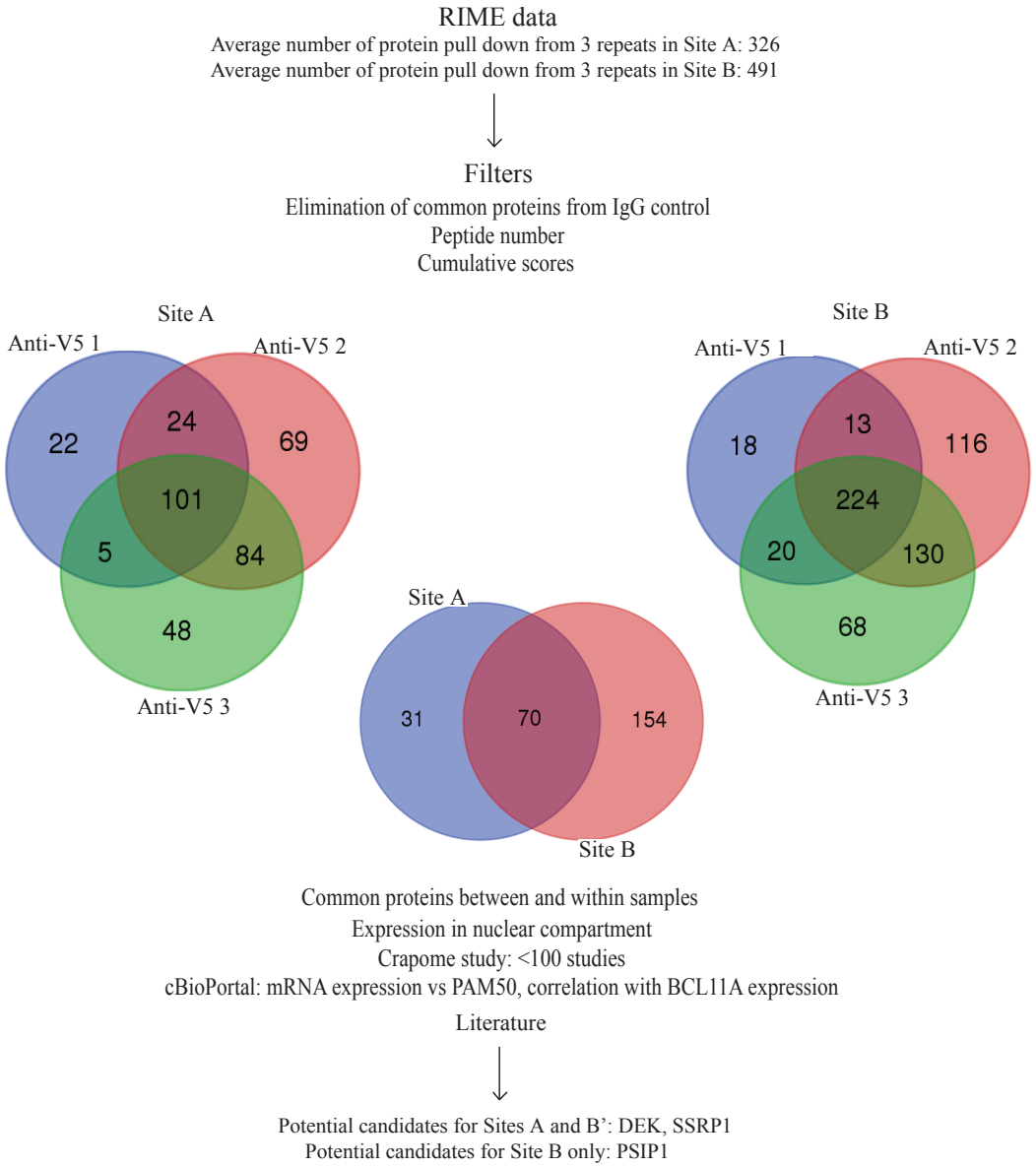


Figure 4.2: Interpreting RIME data.

Average number of proteins pulled down from the triplicates of Sites A and B were 326 and 491, respectively. Proteins were filtered using various strategies – eliminating common proteins from IgG control, number of peptide number plus cumulative scores, common proteins between and within samples, expression in the nucleus, less than 100 studies from the Crapome database, cBioportal database correlating mRNA expression with PAM_{50}, and *BCL11A* expression. literature search for proteins with known roles in breast cancer. 3 potential candidates were identified: DEK and SSRP1 from both target sites and PSIP1 from Site B.

Eliminate from IgG and common between triplicates		
Sites A and B		
UniProt	**Gene ID**	**Gene name**
P21333	FLNA	Filamin-A
O75369	FLNB	Filamin-B
Q14204	DYNC1H1	Cytoplasmic dynein 1 heavy chain 1
Q13838	DDX39B	Spliceosome RNA helicase DDX39B
P38919	EIF4A3	Eukaryotic intiation factor 4A-III
O75533	SF3B1	Splicing factor 3B subunit 1
P13010	XRCC5	X-ray repair cross-complementing protein 5
O15027	SEC16A	Protein transport protein Sec16A
P11940	PABPC1	Polyadenylate-binding protein 1
Q9BY77	POLDIP3	Polymerase delta-interacting protein 3

Table 4.1: Summary of top 20 proteins from the RIME analysis after eliminating common proteins between IgG and triplicates between two target sites.

Top 20 proteins between Sites A and B samples are shown, presented in UniProt ID, gene ID, and gene name.

Eliminate from IgG and common between triplicates		
Site A only		
UniProt	**Gene ID**	**Gene name**
O00567	POLDIP3	Nucleolar protein 56
Q9UKM9	RALY	RNA-binding protein Raly
Q8IY81	GBAA_3676	pre-rRNA processing protein FTSJ3
O60832	DKC1	H/ACA ribonucleoprotein complex subunit DKC1
O43290	SART1	U4/U6.U5 tri-snRNP-associated protein 1
Q9BY78	RNF26	E3 ubiuitin-protein ligase RNF26
Q9Y3Y3	SH2B1	SH2B adapter protein 1
P46014	KAPP	Protein phosphatase 2C 70
Q9Y2X4	SH2D3A	SH2 domain-containing protein 3A
P82980	RBP5	Retinol-binding protein 5

Table 4.2: Summary of top 20 proteins from the RIME analysis after eliminating common proteins between IgG and triplicates at Site A.

Top 20 proteins in Sites A samples are shown, presented in UniProt ID, gene ID, and gene name.

Eliminate from IgG and common between triplicates		
Site B only		
UniProt	**Gene ID**	**Gene name**
Q09666	AHNAK	Neuroblast differentiation-associated protein AHNAK
Q86V81	ALYREF	THO complex subunit 4
O00148	DDX39A	ATP-dependent RNA helicase DDX39A
P68032	ACTC1	Actin, alpha cardiac muscle 1
Q9NZM1	MYOF	Myoferlin
P12956	XRCC6	X-ray repair cross-complementing protein 6
P09651	HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1
P55795	HCRNPH2	Heterogeneous nuclear ribonucleoprotein H2
P23396	RPS3	40S ribosomal protein S3
P17844	DDX5	Probable ATP-dependent RNA helicase DDX5

Table 4.3: Summary of top 20 proteins from the RIME analysis after eliminating common proteins between IgG and triplicates at Site B.

Top 20 proteins in Sites B samples are shown, presented in UniProt ID, gene ID, and gene name.

The Venn diagrams show the number of proteins that overlapped between two or three replicates from the same target site. There were 101 proteins pulled down from Site A and 224 proteins from Site B. The centre of the Venn diagram shows that 70 proteins were identified that were common between the two target sites. It also shows the number of unique proteins from each target site: 31 from Site A and 157 from Site B. The listed proteins were then sorted in descending order by peptide number to identify proteins with a unique peptide number of at least one. Cumulative scores were then pulled from each list. The peptide number of at least one provided confidence that the "correct" protein would be identified from the peptide sequences during mass spectrometry. Each protein was also given a score in direct correlation with the abundance of the proteins in the samples. Thus, the higher the cumulative score, the higher the frequency of detection of that protein during mass spectrometry.

Proteins were also sorted according to their expressions in the cellular compartments. Attention was paid to proteins with expression in the nuclear compartment, because the aim was to study proteins with the potential to regulate expression of *BCL11A*. Tables 4.4 – 4.6 shows the top 20 proteins after the application of this parameter to the process of protein rationalisation. Next, the proteins were considered against the Contaminant Repository of Affinity Purification (Crapome) database [127]. This online resource was used as a quality control to eliminate protein contaminants that had been shown previously to supply noise signals in mass spectrometry studies. This procedure raised the quality of the analysis as the use of negative control is often insufficient to remove those proteins identified as background [127]. Proteins with a score of 100 according to the Crapome database were considered to be noise. The top 20 proteins that remained after this filtering process are shown in Tables 4.7 – 4.9.

The proteins that remained on the list were analysed further against the large dataset of The Cancer Genome Atlas (TCGA) from the cBioPortal [32]. This dataset records the genetic and molecular heterogeneity of breast cancer patients. The parameters employed at this filtering step included comparison of the expression of mRNA against PAM_{50} and its correlation with the expression of *BCL11A*'s mRNA.

Nuclear expression		
Sites A and B		
UniProt	**Gene ID**	**Gene name**
Q9BY77	POLDIP3	Polymerase delta-interacting protein 3
Q9Y3Y2	CHTOP	Chromatin target of PRMT1 protein
P46013	MKI67	Antigen KI-67
P82979	SARNP	SAP domain-containing ribonucleoprotein
Q9Y2X3	NOP58	Nucleolar protein 58
P11388	TOP2A	DNA topoisomerase 2-alpha
Q9Y5S9	RBM8A	RAN-binding protein 8A
Q05519	SRSF11	Serine/arginine-rich splicing factor
Q13185	CBX3	Chromobox protein homolog 3
Q9Y5B9	SUPT16H	FACT complex subunit SPT16

Table 4.4: Summary of top 20 proteins from the RIME analysis after filtering for proteins with nuclear expressions between two targets.

Top 20 proteins in Sites A and B samples are shown, presented in UniProt ID, gene ID, and gene name.

Nuclear expression		
Site A only		
UniProt	**Gene ID**	**Gene name**
O00567	NOP56	Nucleolar protein 56
Q9UKM9	RALY	RNA-binding protein Raly
Q8IY81	FTSJ3	pre-rRNA processing protein FTSJ3
O60832	DKC1	H/ACA ribonucleoprotein complex subunit DKC1
O43290	SART1	U4/U6.U5 tri-snRNP-associated protein 1
Q9BY78	RNF26	E3 ubiuitin-protein ligase RNF26
Q9Y3Y3	SH2B1	SH2B adapter protein 1
P46014	KAPP	Protein phosphatase 2C 70
Q9Y2X4	SH2D3A	SH2 domain-containing protein 3A
P82980	RBP5	Retinol-binding protein 5

Table 4.5: Summary of top 20 proteins from the RIME analysis after filtering for proteins with nuclear expressions at Site A.

Top 20 proteins in Site A samples are shown, presented in UniProt ID, gene ID, and gene name.

Nuclear expression		
Site B only		
UniProt	**Gene ID**	**Gene name**
Q09666	AHNAK	Neuroblast differentiation-associated protein AHNAK
Q9NZM1	MYOF	Myoferlin
P62333	PSMC6	26S proteasome regulatory subunit 10B
Q96S55	WRNIP1	ATPase WRNIP1
Q8ND24	RNF214	RING finger protein 214
O75367	H2AFY	Core histone macro-H2A.1
P04083	ANXA1	Annexin A1
O00299	CLIC1	Chloride intracellular channel protein 1
Q86V48	LUZP1	Leucine zipper protein 1
Q9H0S4	DDX47	Probable ATP-dependent RNA helicase DDX47

Table 4.6: Summary of top 20 proteins from the RIME analysis after filtering for proteins with nuclear expressions at Site B.

Top 20 proteins in Site B samples are shown, presented in UniProt ID, gene ID, and gene name.

Crapome database (<100 studies)		
Sites A and B		
UniPot	**Gene ID**	**Gene name**
Q9Y3Y2	Chtop	Chromatin target of PRMT1 protein
P46013	MKI67	Antigen KI-67
Q9Y2X3	NOP58	Nucleolar protein 58
P11388	TOP2A	DNA topoisomerase 2-alpha
Q9Y5S9	RBM8A	RNA-binding protein 8A
Q05519	SRSF11	Serine/arginine-rich splicing factor 11
Q13185	CBX3	Chromobox protein homolog 3
Q9Y5B9	SUPT16H	FACT complex subunit SPT166
Q08945	SSRP1	FACT complex subunit SSRP1
P35659	DEK	Protein DEK

Table 4.7: Summary of top 20 proteins from the RIME analysis after referring to the Crapome database between Sites A and B.

Top 20 proteins in Sites A and B samples are shown, presented in UniProt ID, gene ID, and gene name.

Crapome database (<100 studies)		
Site A only		
UniProt	**Gene ID**	**Gene name**
O00567	NOP56	Nucleolar protein 56
Q9UKM9	RALY	RNA-binding protein Raly
Q8IY81	FTSJ3	pre-rRNA processing protein FTSJ3
O60832	DKC1	H/ACA ribonucleoprotein complex subunit DKC1
O43290	SART1	U4/U6.U5 tri-snRNP-associated protein 1
Q9BY78	RNF26	E3 ubiquitin-protein ligase RNF26
Q9Y3Y3	SH2B1	SH2B adapter protein 1
P46014	KAPP	Protein phosphatase 2C 70
Q9Y2X4	SH2D3A	SH2 domain-containing protein 3A
O75475	PSIP1	PC4 and SFRS-1 interacting protein

Table 4.8: Summary of top 20 proteins from the RIME analysis after referring to the Crapome database at Site A.

Top 20 proteins in Site A samples are shown, presented in UniProt ID, gene ID, and gene name.

Crapome database (<100 studies)		
Site B only		
UniProt	**Gene ID**	**Gene name**
Q09666	AHNAK	Neuroblast differentiation-associated protein AHNAK
Q9NZM1	MYOF	Myoferlin
Q9BY77	POLDIP3	Polymerase delta-interacting protein 3
P62333	PSMC6	26S proteasome regulatory subunit 10B
Q96S55	WRNIP1	ATPase WRNIP1
P82979	SARNP	SAP domain-containing ribonucleoprotein
Q8ND24	RNF214	RING finger protein 214
P04083	ANXA1	Annexin A1
O00299	CLIC1	Chloride intracellular channel protein 1
Q86V48	LUZP1	Leucine zipper protein 1

Table 4.9: Summary of top 20 proteins from the RIME analysis after referring to the Crapome database at Site B.

Top 20 proteins in Site B samples are shown, presented in UniProt ID, gene ID, and gene name.

PAM_{50} is an assay that classifies intrinsic subtypes through the quantification of the co-expression of 50 genes [27]. Only proteins that were found to be highly regulated in TNBC were advanced to the next filtering stage. Application of all these parameters led to identification of 12 potential protein candidates. The plots of their mRNA expressions against PAM_{50} are shown in Figure 4.3.

Finally, a literature review was performed to search for proteins with known functions in the regulation of transcription or involvement in breast cancer. Three proteins were identified from Sites A and B after filtering the lists of RIME pull-down proteins against these multiple parameters and resources. These proteins were DEK, structure specific recognition protein 1 (SSRP1) and PC4 and SFRS1 interacting protein 1 (PSIP1).

4.2.3 The validation of protein pull-down from RIME

4.2.3.1 The expression of candidates in TNBC cell lines

To confirm whether DEK, SSRP1 and PSIP1 were abundantly expressed in TNBC, protein lysate was extracted from a panel of breast cancer cell lines and their quantities determined. Seven TNBC and three Luminal lines were used. They were studied by western blot analysis for which conditions for the employed antibodies had been optimised. The optimisations were performed with HEK293 cells to choose the dilutants for the primary antibodies that would retain consistent concentrations. Chapter 2: Materials and Methods provides further details.

DEK, SSRP1 and PSIP1 were all observed to show greater expression in the TNBC cell lines than in the Luminal cell lines. The western blots obtained can be seen in Figure 4.4. Tubulin was used as a loading control.

4.2.3.2 Confirmation by ChIP that the proteins bind at the target sites

ChIP experiments were performed in MDA-MB-231 wild type cells to determine whether DEK, SSRP1 and PISP1 bind at Sites A and B of *BCL11A*. Antibodies specific to each candidate or to the IgG control were used to pull-down proteins.

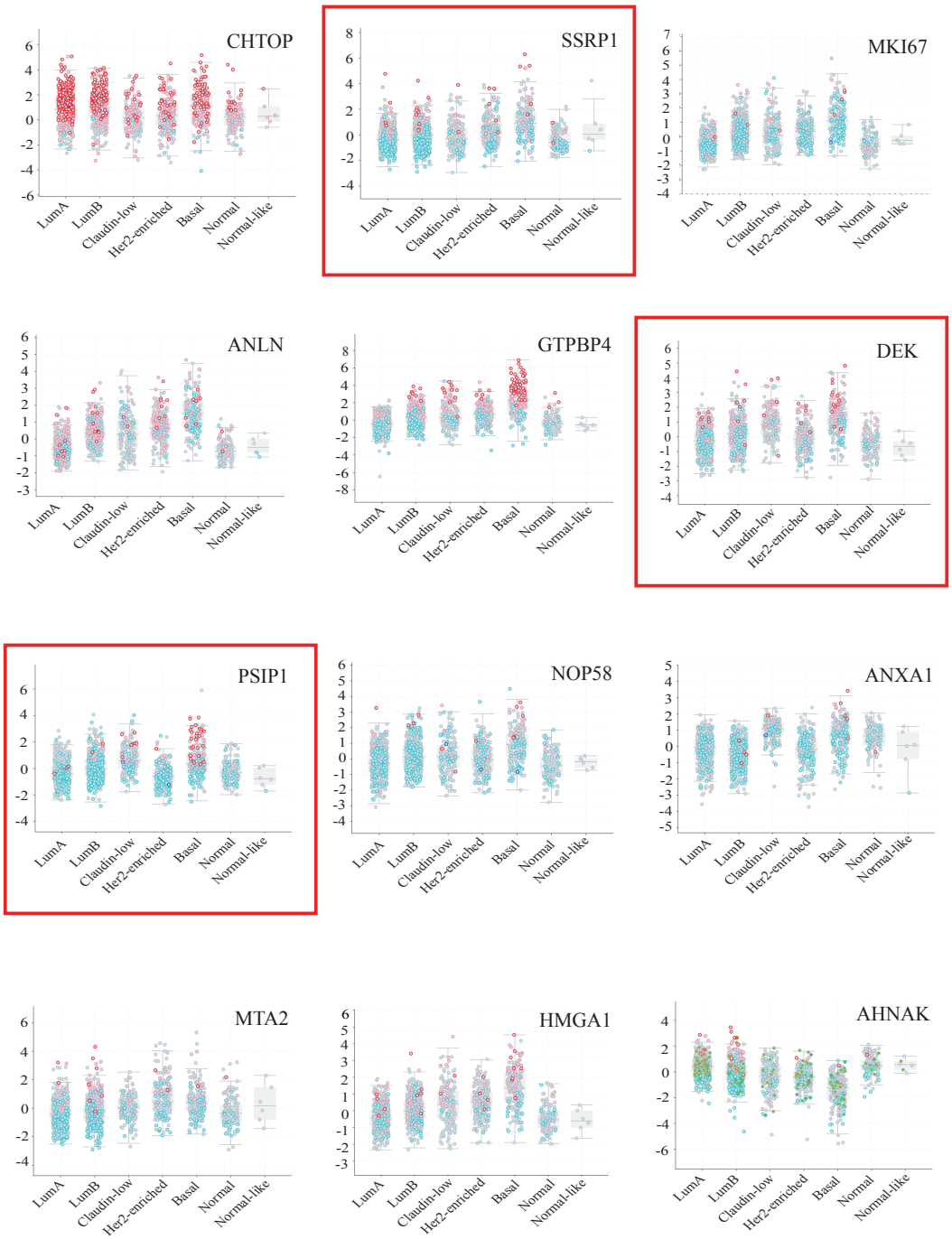


Figure 4.3: cBioportal database correlating mRNA expression with PAM_{50} and *BCL11A* expression.

Graphs were adapted from the cBioportal database; proteins CHTOP, SSRP1, MKI67, ANLN, GTPBP4, DEK, PSIP1, NOP58, ANXA1, MTA2, HMGA1, and AHNAK all match the above criteria.

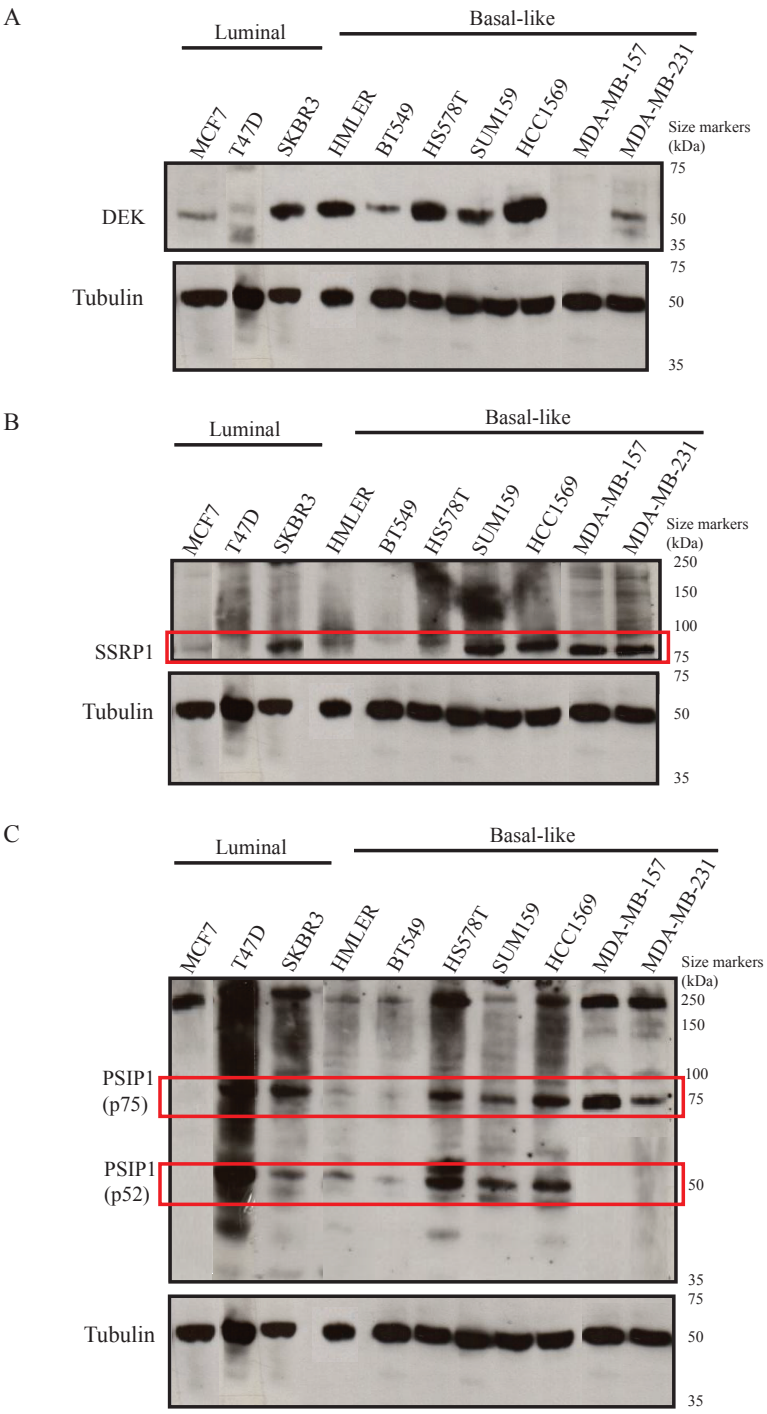


Figure 4.4: Protein expression of candidates on a panel of breast cancer cell lines.

A panel of breast cancer cell lines were used. DEK expression is found in all cells except MDA-MB-157. SSRP1's varying expressions were found across the panel. Different concentrations of PSIP1 expression were found across the panel of cells.

Primers were then used to amplify the targets in the qPCR step to identify the bound DNA regions. The ChIP analysis is shown in Figures 4.5. It indicated that DEK, SSRP1 and PSIP1 binding was enriched at Sites A and B. Moreover, binding by SSRP1 at Site B and by PSIP1 at both target sites reached statistical significance.

The specificities of protein binding were then assessed by repeating the ChIP experiments. Proteins were pulled down as described earlier but with the use of different primers during the qPCR step. These primers, which were labelled primers 3 and 4, were designed to be placed as close as possible to Sites A and B. Their positions on the *BCL11A* gene can be seen in Figure 4.6 A.

The ChIP data displayed in Figures 4.6 B-D show that the binding of DEK, SSRP1 and PSIP1 was enriched at neighbouring sites to A and B. Moreover, this binding reached statistical significance in the case of SSRP1 at the region next to Site A and for PSIP1 at the region adjacent to Site B.

These ChIP data were also validated in another TNBC cell line, HCC1569. The western blot data shown in Figure 4.4 indicated that this cell line had more DEK, SSRP1 and PSIP1 expressions than were observed in MDA-MB-231 cells. As before, antibodies specific to the candidate proteins or to the IgG control were used to pull-down proteins and then their binding was analysed to determine whether they were bound at Sites A and B or at neighbouring regions. The data shown in Figure 4.7 suggests that binding of DEK, SSRP1 and PSIP1 was enriched at Sites A and B. The binding by DEK and SSRP1 at Site A reached statistical significance.

Finally, an evaluation was performed regarding the specificity of the binding. Figure 4.8 shows that DEK, SSRP1 and PSIP1 were bound at sites that lay near A and B.

To summarise: DEK, SSRP1 and PSIP1 bind at Sites A and B and at neighbouring sites.

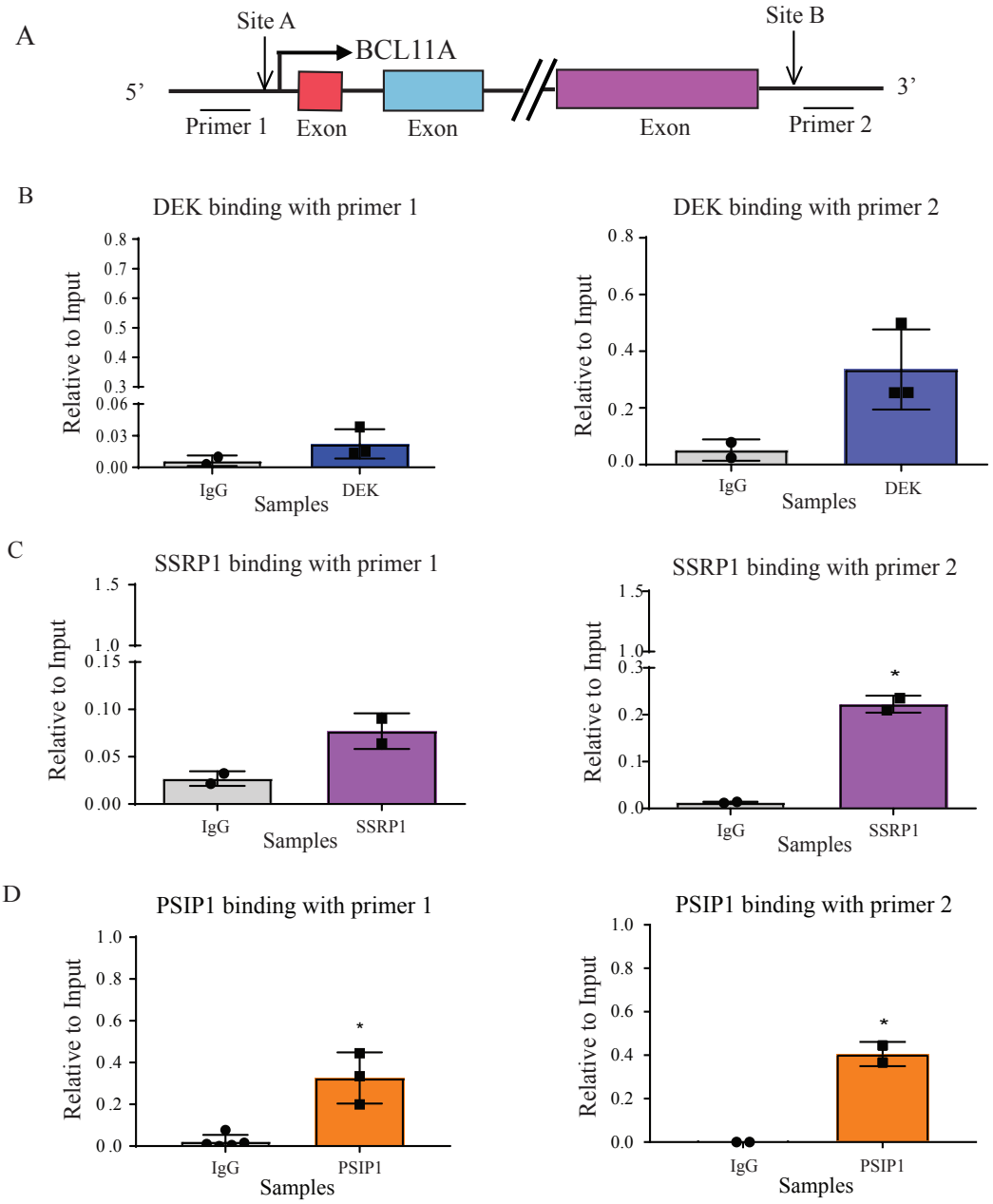


Figure 4.5: ChIP experiment results, showing candidates binding in MDA-MB-231.

(A) in Figure 4.5 above is a graphic representation of the positions of primers 1 and 2 on *BCL11A*. DEK, SSRP1, and PSIP1 bind at both targets but SSRP1 binds preferentially at or close to Site B. Samples pulled down with IgG antibody were used as internal controls and graphs are plotted in relation to input. (*<0.05, t-test, n=2-3.)

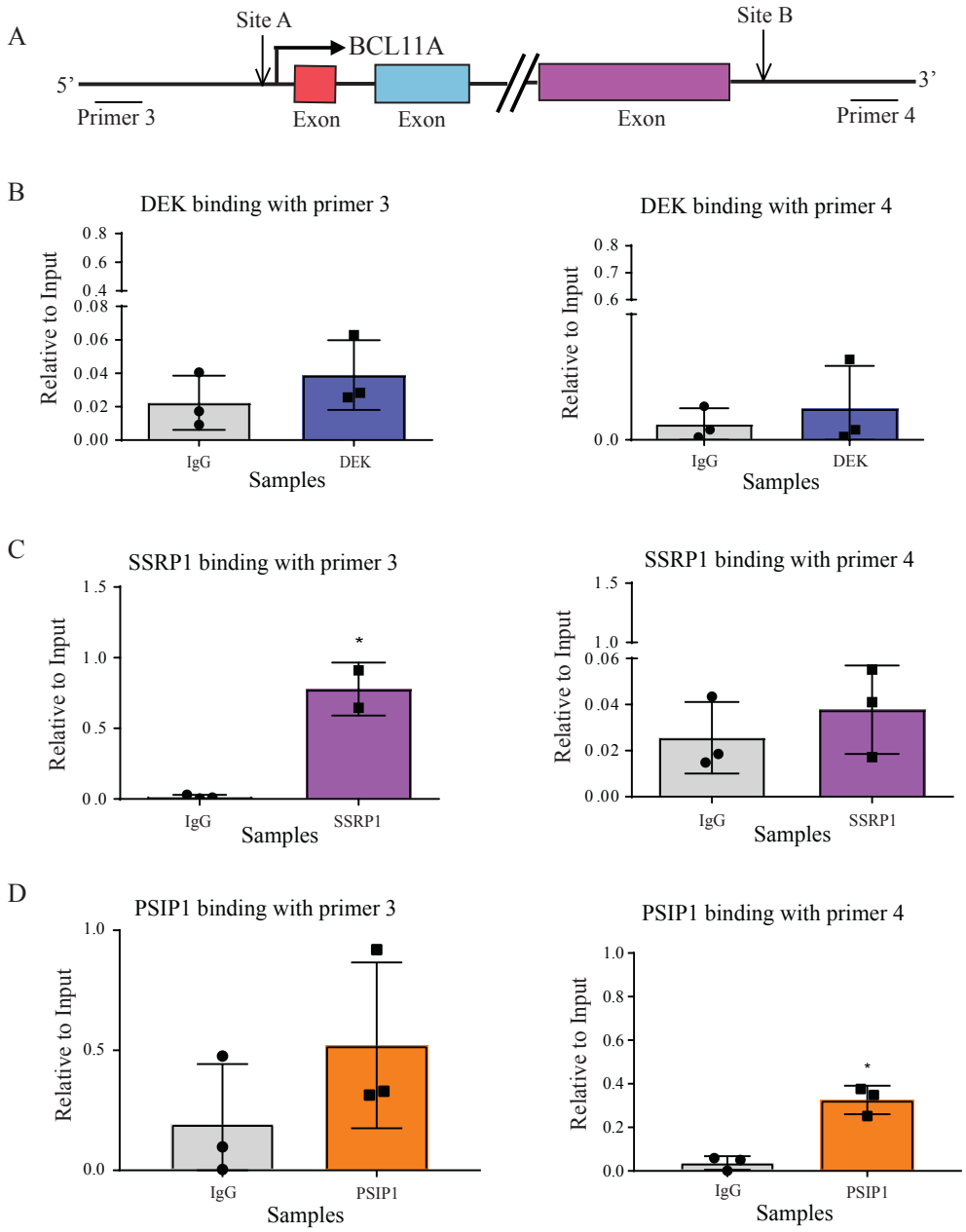


Figure 4.6: Specificities of candidates' bindings in MDA-MB-231 cells.

(A) in Figure 4.6 above is a graphical representation of the positions of primers 3 and 4 on *BCL11A*. These are nearby regions of Sites A and B. DEK, SSRP1, and PSIP1 bind to both targets, but SSRP1 was observed to bind preferentially to regions close to Site A whereas PSIP1 was observed to bind preferentially to Site B. Samples pulled down with IgG antibody were used as internal controls and graphs are plotted in relation to input. (* <0.05, t-test, n=2-3.)

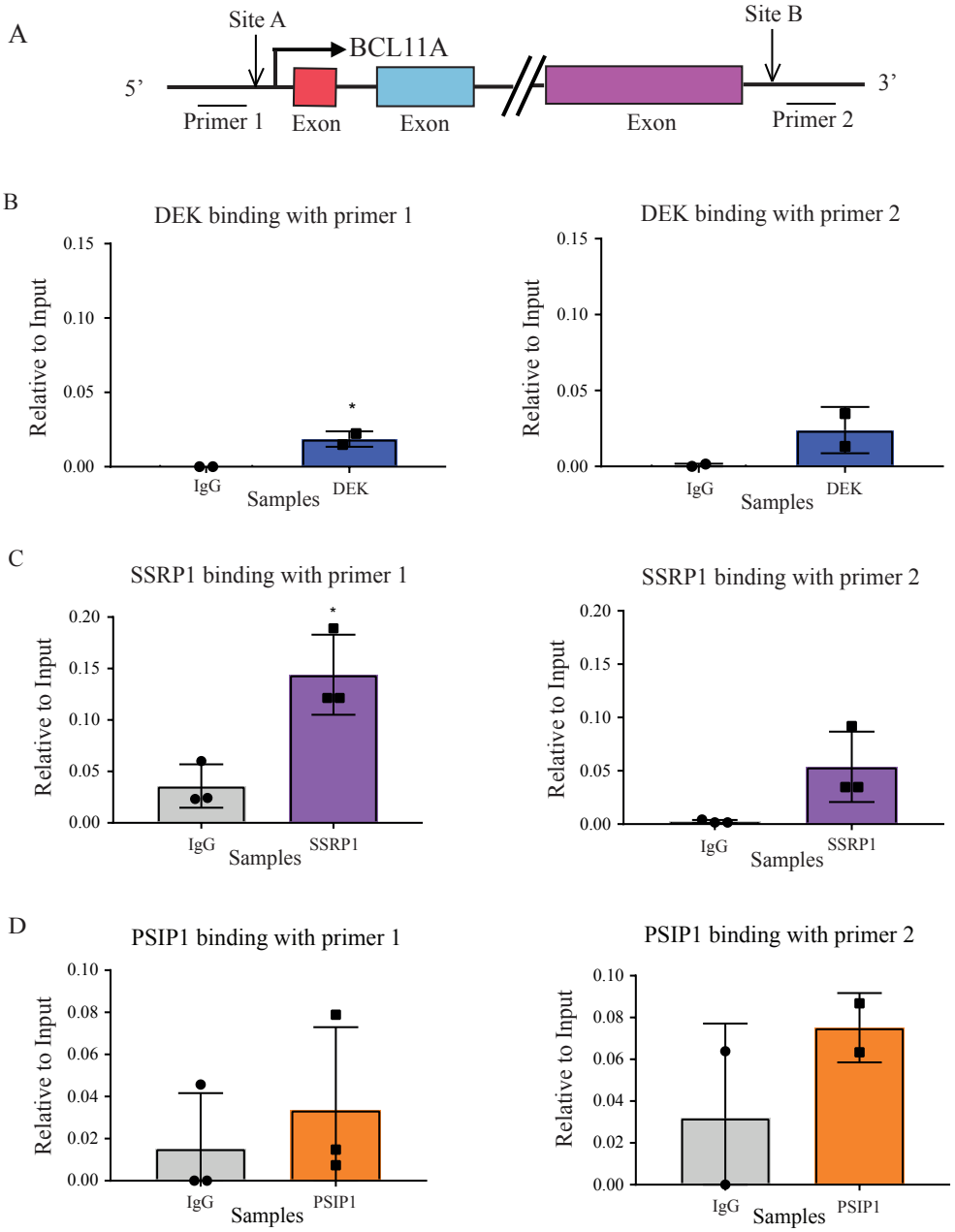


Figure 4.7: ChIP experiment results, showing candidate bindings in HCC1569 cells.

(A) in Figure 4.7 above is a graphical representation of the positions of primers 1 and 2 on *BCL11A*. All candidate proteins were observed to bind to both target sites, but SSRP1 bound preferentially at the 5' end. Samples pulled down with IgG antibody were used as internal controls and graphs are plotted in relation to input. (*<0.05, t-test, n=2-3.)

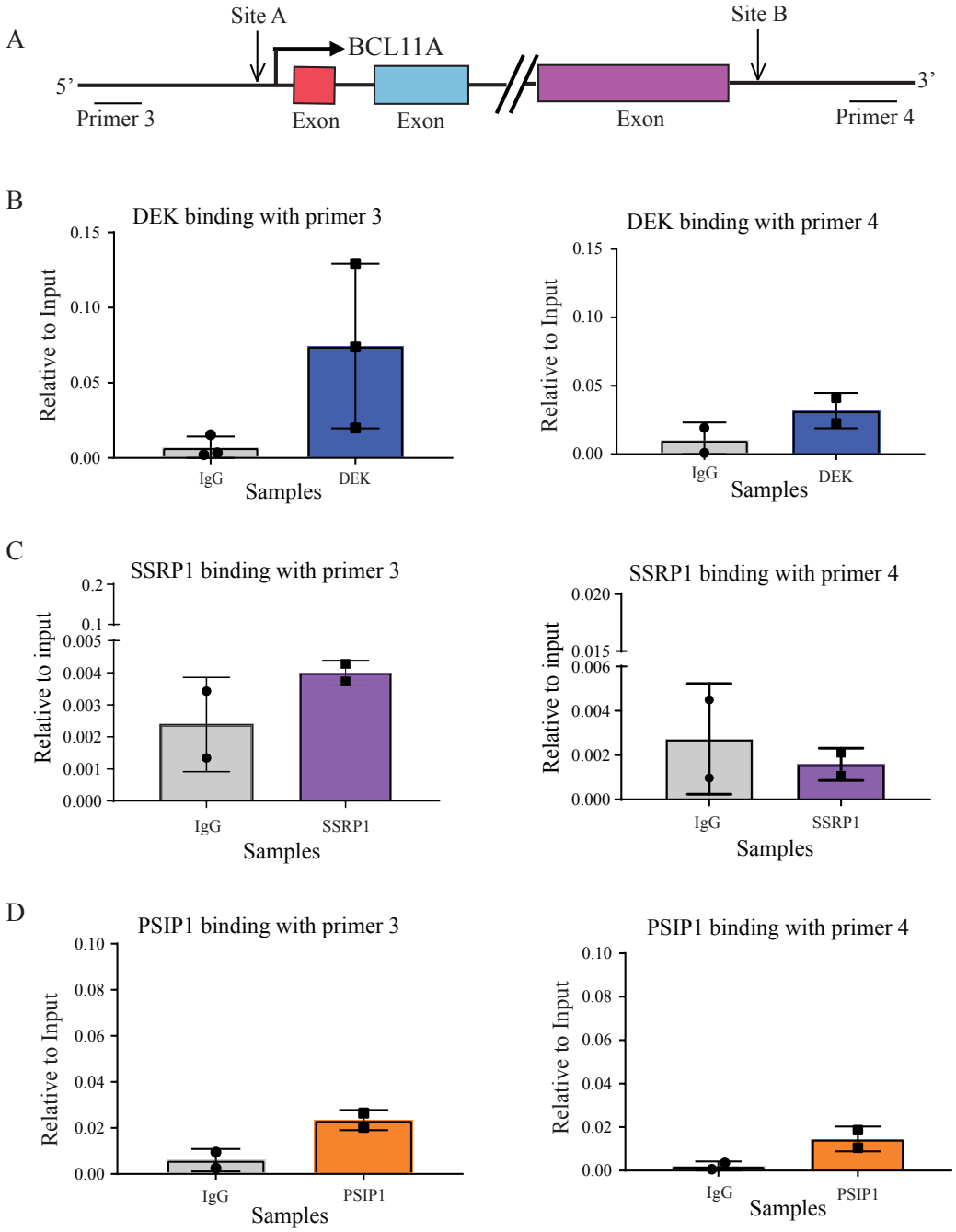


Figure 4.8: Specificities of candidate bindings in HCC1569 cells.

(A)in Figure 4.8 above is a graphical representation of the positions of primers 3 and 4 on *BCL11A*. These are the neighbouring regions of Sites A and B. DEK, SSRP1, and PSIP1 were observed to bind at both target sites. Samples pulled down with IgG antibody were used as internal controls and graphs are plotted in relation to input. (*<0.05, t-test, n=2-3.)

4.2.3.3. The evaluation of any bias in target binding by DEK, SSRP1 and PSIP1

To assess whether DEK, SSRP1 or PSIP1 show binding bias regarding particular sites on *BCL11A* in MDA-MB-231 and HCC1569 cell lines, statistical analysis was performed on the results described in section 4.2.3.2. The potential for bias was tested between Primers 1 and 2 at Sites A and B; between Primers 3 and 4 at the regions near to Sites A and B; and among all four primers. The results are presented in Table 4.10.

Firstly, DEK binding was found to differ between Sites A and B in MDA-MB-231 cells. This difference was also observed when results were compared across all four primers and was therefore further analysed with post-hoc tests. It was revealed that Primer 1, at Site A, bound differently to all other primers. Primer 2 was also found to bind differently at Site B compared with Primers 3 and 4 at their binding sites.

Secondly, SSRP1 was found to bind in a similar way to DEK, except that differences in binding were also observed between Primers 3 and 4. Post-hoc tests showed statistically significant binding differences between Primers 1 and 2 compared with Primers 3 and 4. In contrast, no differences were found between methods of PSIP1 binding at any of the sites studied.

DEK binding bias was observed to be different in HCC1569 compared with that in MDA-MB-231 cells, in that no statistically significant binding difference was detected in this cell line between any primers. In contrast, SSRP1 binding differences between Primers 1 and 2 were statistically significant, and subsequent tests that analysed all four primers also suggested binding differences among them. Lastly, the binding bias of PSIP1 was found to be the same as in MDA-MB-231 cells in that the differences between target sites and nearby regions were not statistically significant.

Proteins	MDA-MB-231 cells			HCC1569 cells		
	Primers 1 vs 2	**Primers 3 vs 4**	**Primers 1-4**	**Primers 1 vs 2**	**Primers 3 vs 4**	**Primers 1-4**
DEK	Significant	Not significant	Significant	Not significant	Not significant	Not significant
SSRP1	Significant	Significant	Significant	Significant	Not significant	Significant
PSIP1	Not significant	Not significant	Not significant	Not significant	Not significant	Not significant

Table 4.10: Summary of statistical analysis of binding bias by DEK, SSRP1 and PSIP1 in MDA-MB-231 and HCC1569 cell lines.

Statistical analysis of protein bindings was performed at the two target sites, Sites A and B (Primers 1 and 2), and their nearby regions (Primers 3 and 4). DEK bindings showed statistical significance at Sites A and B and between the four primers in MDA-MB-231 cells but not in HCC1569. SSRP1 bindings were significantly different across all sites in MDA-MB-231 but not when comparing Primers 3 and 4 in the HCC1569. Lastly, PSIP1 bindings show no significant difference across both cell lines.

4.3 Discussion

After the generation of a stable cell line with gRNA- and dCas9-expressing plasmids, the numbers of cells were expanded and fixed by formaldehyde. Next, sonication was employed for optimisation, followed by immunoprecipitation using antibody against the multi-tagged dCas9 to pull-down dCas9 itself and proteins in close proximity. Samples were then analysed by mass spectrometry. The positive protein findings identified by RIME were based on various criteria, including expression of the proteins in the cellular compartments, similarity to *BCL11A* in terms of clinical contribution, and functional properties. Three proteins, DEK, SSRP1 and PSIP1, were found to exhibit relationships to the clinical characteristics of Triple-negative breast cancer and tumour progression, and some were involved in the maintenance of stemness. These findings proved that pulling down proteins in relation to breast cancer was feasible through the use of the experimental design employed in this study.

4.3.1 Analysis of RIME pull-down

Pulled-down proteins have various functions and roles in breast cancer, which are summarised in Table 4.11. Their relationship with clinical features and involvement in certain processes suggested that the experimental design employed in this study was applicable to the investigation of potential protein regulators of *BCL11A* transcription.

The proteins were first split into peptides with the use of trypsin to perform mass spectrometry. Three peptide filters were employed to increase the possibility of Correct protein identification. Secondly, the data was rationalised through the elimination of proteins against the IgG control and the identification of unique proteins present in each sample. Thirdly, the proteins were separated into different cellular compartments and only those observed to be from the nucleus were further analysed. This was because the focus of the project was to identify putative protein regulators for *BCL11A*. Next, proteins that were expressed in the nucleus were analysed using the Crapome database, which provides hints regarding the

	DEK	SSRP1	PSIP1
Functions	Induce positive DNA supercoils [128], chromatin organisation [129] and selection of position for splicing [130]	Subunit of the FACT complex and is required for transcription [131], as co-activator of p63/TP63 [132] and stem cell transcription factors [133]	Involved in lens epithelial cell differentiation [134], as transcriptional co-activator for neuroepithelial stem cells [135] and lentiviral integration [136]
Roles in breast cancer	Prognostic marker for survival [137]; involves in metastasis [138] [139], mammosphere formation [138] and maintaining stemness [138] [139]	Correlates with negative hormonal receptor status, aggressiveness and poor prognosis [140] [141] [142]; involves in cell proliferation, transformation and maintaining stemness [142] [143]	Negative correlation to survival, regulates cell cycle gene expression , involves in migration and invasion [144]

Table 4.11: Summary of known cellular functions and roles in breast cancer. These factors identified from RIME are involved in different roles in the cells. Some were also identified to be useful as prognostic markers and are involved in different functions in breast cancer.

background proteins that are commonly identified from mass spectrometry experiments [127]. Comparison with this data enabled further rationalisation to continue the study only with those proteins that showed higher potential for success related to the project aim. Lastly, proteins were identified according to their clinical contributions to breast cancer subtypes. This was achieved using the PAM_{50} analysis from the TCGA dataset. Application of these filters identified proteins DEK, SSRP1 and PSIP1, which were seen only in samples pulled down using Anti-V5 antibody with expressions in the nucleus and upregulated in the TNBC subtype. These proteins were analysed further based on their cellular functions.

To investigate whether DEK, SSRP1 and PSIP1 showed greater expression in cell lines, as stated in the cBioportal database, a panel of breast cancer cell lines was collected and proteins lysate were extracted to perform western blot analysis.

DEK was originally identified as a gene involved in acute myeloid leukaemia (AML) [145]; in chromatin organisation [129]; in tumour progression [138] [139]; in the maintenance of stemness [138] [139]; in the formation of mammosphere in breast cancer [138]; and it was found to be expressed in proliferating cells [146]. The western blot displayed in Figure 4.4 showed that the expression of DEK was upregulated in TNBC cell lines. However, its expression was also observed across the Luminal cell lines. One explanation could be that it is important for cell maintenance, given that its normal cellular functions include the regulation of transcription, the splicing of mRNA and the replication and repair of DNA [128] [129] [130].

The expression of SSRP1 was also found to be upregulated in TNBC cell lines. The western blot shown in Figure 4.4 B also shows that there was strong expression in some Luminal cell lines. This could be because SSRP1 plays a general role as a chromatin-associated protein, forming a FACT complex with subunit FACT complex subunit SPT16 (SUPT16) [131]. It is also involved in: DNA replication and repair [147] [148], the disassociation of histone proteins to regulate gene transcription [140], and as a transcriptional co-activator for p63 [132]. It has also

been found to promote cell proliferation, and transform and maintain cell stemness in breast cancer [142] [143].

Lastly, the expression of PSIP1 was determined. It was found to be upregulated in the TNBC cell lines. However, as with SSRP1, there were also PSIP1 expressions in some Luminal cell lines. This could be because it plays an important role in general cell function. PSIP1 was originally identified as taking part in the regulation of transcription in lens epithelial cells [134]. It also acts as a transcription co-activator [135], as a promoter of cell proliferation [134] and as a repairer of DNA double-strand breaks [149]. These general roles are important for cell survival and could explain its expression in both TNBC and Luminal subtypes. Furthermore, it is known that PSIP1 regulates the expression of genes in the cell cycle, and promotes migration and invasion in breast cancer [144]. Table 4.11 summarises the known functions of DEK, SSRP1 and PSIP1 in breast cancer.

4.3.2 Evaluation of the binding of candidates in TNBC cell lines

Antibodies specific to DEK, SSRP1 and PSIP1 were used in ChIP studies to determine whether they were true proteins that bound to Sites A and B on *BCL11A* in MDA-MB-231 cells. The experiments showed that all of their bindings were enriched at both target sites, with some being statistically significant. Next, their binding specificities to these sites were evaluated through the use of primers which amplified regions that flanked Sites A and B. These primers were designed to be as close to the targets as possible, and thus it was not surprising to find enrichment in these regions as well.

The findings were also confirmed in another independent cell line, HCC1569. This cell line was shown through the use of western blots to exhibit stronger expressions of DEK, SSRP1 and PSIP1 than in MDA-MB-231 cells. The results are shown in Figure 4.4. The ChIP data gathered from this cell line led to a similar conclusion to that drawn from the work with MDA-MB-231 cells, which was that DEK, SSRP1 and PSIP1 bindings were enriched at Sites A and B, some of which enrichment was again statistically significant. As described earlier, the specificities of their bindings

were also assessed using primers specific to nearby regions. The bindings of all three proteins were enriched at the nearby sites.

In summary, the results for these two TNBC cell lines show that the bindings of DEK, SSRP1 and PSIP1 were enriched at the target sites and at their neighbouring regions.

We also wanted to assess whether DEK, SSRP1 and PSIP1 showed binding bias towards particular sites, and this was accomplished by way of statistical analysis on the ChIP results. We can conclude that this potential binding bias depends on the cell lines tested. For instance, DEK binding did not display statistically significant differences when comparing primers against each other in HCC1569, while the same cannot be said for MDA-MB-231 cells. The results also suggest that SSRP1 showed binding bias towards the sites studied in MDA-MB-231 cells, but this bias was shown towards only two of these sites in HCC1569. PSIP1 bindings were found to show no statistically significant binding bias in either cell line.

To summarise, DEK bindings were enriched at Sites A and B in both cell lines but the bias of its binding towards specific sites was inconclusive. One reason could be that the sample size used in the experiment was too small for post-hoc tests to detect robust statistical significance in observed differences. SSRP1 bindings were also enriched at the target sites and statistical analysis showed some bias towards particular binding sites. Finally, PSIP1 bindings were found to be enriched at both target sites without any bias.

4.3.3 Do DEK, SSRP1 and PSIP1 form a complex?

Analysis of the Search Tool for the Retrieval of Interacting Genes (STRING) database has suggested that these three proteins might interact together, with DEK as a central protein in human cells. However, no recognised interactions with *BCL11A* were found. This STRING database is an online resource with known and predicted protein-protein interactions sourced from experimental data and computational assessments [150].

Co-immunoprecipitation experiments were performed to investigate whether true interactions were present. Magnetic beads were incubated with DEK antibody and attempts were made to pull-down SSRP1 and PSIP1 from MDA-MB-231 cells. Initial attempts with this cell line proved to be technically difficult. To overcome this, the experiments were performed in HCC1569 cells. This was due to the greater expression of DEK, SSRP1 and PSIP1 in this cell line in comparison with MDA-MB-231 cells. However, due to time constraints, the protocol was not optimised. Matysiak's study has previously shown that SSRP1 interacts with the PSIP1 p75 isoform to regulate HIV-1 integration [151]. Therefore, investigation of a possible similar interaction in TNBC would be worthwhile. Co-IP experiments will be followed up in the future.

In conclusion, DEK, SSRP1 and PSIP1 were identified through the RIME analysis to be putative regulators of *BCL11A*. Their bindings and ability to form a complex were also investigated. Chapter 5 will describe their functions and impact on the expression of *BCL11A*.

CHAPTER 5

Investigating the functional effects of DEK, SSRP1 and PSIP1 using shRNA knock-down

5.1 Introduction

DEK, SSRP1 and PSIP1 were identified through the analysis of the RIME proteomics, as discussed in Chapter 4. These factors have known roles in breast cancer, which are summarised in Table 4.11.

Apart from the roles already identified that DEK plays in breast cancer, one study has shown that translation of DEK is dependent on transfer RNA (tRNA) post-transcriptional modifying enzymes, elongator complex protein 3 (ELP3) and thiouridylase proteins 1/2 (CTU1/2). It has also shown that DEK facilitates the internal ribosome entry site-dependent (IRES-dependent) translation of lymphoid-enhancer binding factor 1 (LEF1) to promote breast cancer invasion and metastasis [152]. This suggests that DEK may have a role at both transcription and translation level in breast cancer.

On the other hand, the complex that facilitates chromatin transcription, or the FACT complex, in which SSRP1 is a subunit, has been found to bind to DNA and block replication or repair during cisplatin treatment in non-small-cell lung cancer. Hence, SSRP1 plays a part in an important mechanism which facilitates cisplatin-induced apoptosis [153].

To investigate the effects of DEK, SSRP1 and PSIP1, and in particular their impact on the expression of *BCL11A* in TNBC, expression of these proteins was knocked down using short hairpin RNA (shRNA). These shRNA molecules bind to their target mRNA transcript through complementary base pairing, which leads to degradation and thus a reduction in expression [112].

The aim of the work described in this chapter is to investigate the effects of DEK, SSRP1 and PSIP1 through the knocking-down of their expressions with shRNA to determine whether these proteins participate in the regulation of the expression of *BCL11A* in TNBC.

5.2 Results

5.2.1 The design of primers for the shRNA-expression plasmid

In order to knock-down the expression of DEK, SSRP1 and PSIP1, sequences of shRNA primers were first designed and obtained from the MISSION® Sigma database. This database contains verified shRNA sequences that target specific regions of the mRNA transcript, and users can design shRNA primers for their genes of interest [118]. These sequences along with their complementary sequences were then blasted against the human genes for DEK, SSRP1 and PSIP1 from the UCSC genome browser to verify their specificities regarding target regions. The position of the targets covered through the use of these shRNA primers are shown in Figure 5.1 A. Scrambled shRNA, which does not target any known human or mouse genes, was used as a control in the following experiments.

5.2.2 The generation of stable cell lines that express shRNA

After the sequences were obtained and verified, these shRNA primers were cloned into the expression plasmid that expresses a geneticin-resistance gene and EGFP fluorescence signals. The features of the shRNA-expression plasmid are shown in Figure 5.1 A.

The shRNA-expression plasmids were then transfected into MDA-MB-231 cells. The geneticin-resistance gene enabled the selection of transfected cells and generation of a stable cell line by FACS for EGFP-positive cells. The schematic of this experimental design is shown in Figure 5.1 B. In addition to these selection markers, the plasmid also adopts the PB transposon system through a cut-and-paste mechanism, as described in Chapter 3. After the establishment of cell lines which constitutively expressed these shRNA plasmids, the degree of knock-down of their target genes would be determined and their functional effect investigated.

5.2.3 The validation of shRNA knock-down

The first validation performed after the transfection of these primers into MDA-

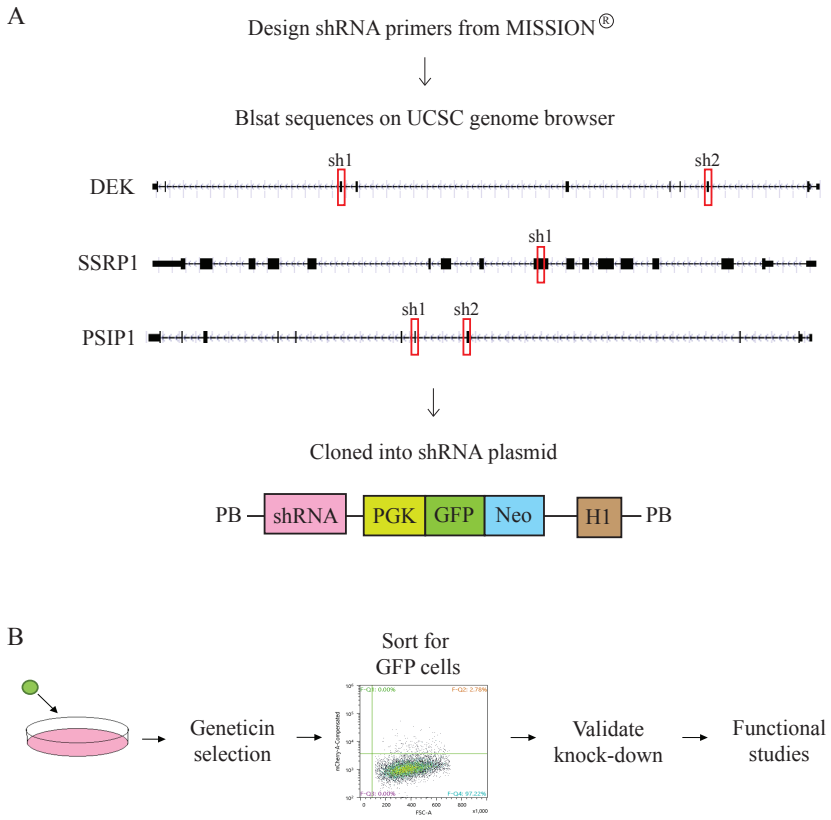


Figure 5.1: The design of shRNA primers and the schematic of the experimental plan.

(A) shRNA primers for DEK, SSRP1 and PSIP1 were first obtained from MISSION, Sigma, and blasted against the corresponding genes taken from the UCSC genome browser. The red boxes show the positions of shRNA targeting on each gene. There were two targets for DEK and PSIP1, and one for SSRP1. Next, these primers were cloned into the shRNA-expression plasmid with the selection markers EGFP and the geneticin-resistance gene through an adaptation of the PB transposon system. (B) shows the schematic for the generation of stable cell lines. The shRNA-expressing plasmids were transfected into MDA-MB-231 cells, followed by geneticin selection and FACS. Transfected cells were collected after the knock-down of their targets by shRNA and their functional effects were then investigated.

MB-231 cells was to determine whether the expressions of their targets, DEK, SSRP1 and PSIP1, were knocked down by shRNA. Transfected cells were collected, RNA extracted, and expression investigated using qPCR. One primer per gene was used in the qPCR step and the expressions were plotted on a graph in relation to the fold change of the control shRNA (CSH). The CSH plasmid contained scrambled shRNA as mentioned in section 5.2.1.

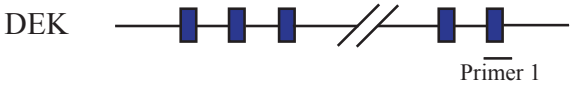
The position of the primer used to detect DEK expression is shown in Figure 5.2 A and qPCR data are listed in 5.2 B. The result showed successful knock-down of DEK expression in both shRNAs, and more knock-down was achieved when shRNA 1 was used.

The expression of SSRP1 was studied in the same way. One primer was chosen; its position is shown in Figure 5.3 A. The qPCR plot displayed in Figure 5.3 B showed that the expression of SSRP1 was knocked down by more than half in comparison with the CSH sample.

Lastly, qPCR was performed to quantify the expression of PSIP1. The position of primer used is shown in Figure 5.4 A. The data displayed in Figure 5.4 B showed successful knock-down when treated with both shRNAs, although shRNA 1 was shown to be more efficient at knock-down than shRNA 2.

The knock-down of factors through the use of shRNA was also investigated at a protein level by collecting protein lysates and performing western blots, as shown in Figure 5.5. CSH was used as a control and tubulin as a loading control. It can be seen that the knock-down of DEK expression was reflected at the protein level. However, SSRP1 knock-down was not reflected when SSRP1 shRNA 1 plasmid was employed, as seen in the western blot. The antibody used to perform this western blot detected both isoforms of PSIP1, p75 and p52. The shRNA 1 sample was seen to contain less p75 isoform than the CSH and shRNA 2 samples. However, no difference was observed in the quantity of p52 isoform between the PSIP1 shRNAs and CSH.

A



B

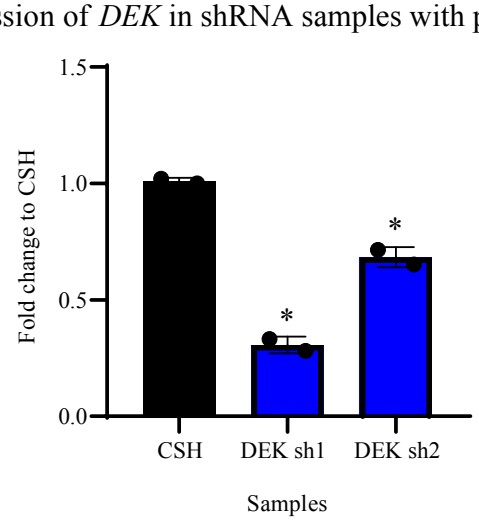
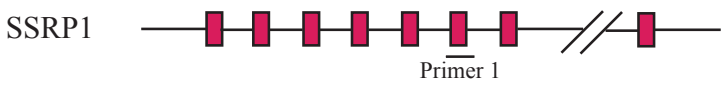


Figure 5.2: Validation of the expression of *DEK* gene knock-down with RT-qPCR.

(A) shows the positions on *DEK* of primer 1 used in the qPCR step. The blue boxes represent the exons. (B) shows the graph of *DEK* gene knock-down, plotted in relation to the fold change to CSH control. *DEK* expression was knocked-down in the DEK shRNA 1 sample to a greater degree than in the DEK shRNA 2 sample. (* <0.05, One-way Anova, Dunnett post-hoc test, n=2 (Average value of two sets each performed in triplicate))

A



B

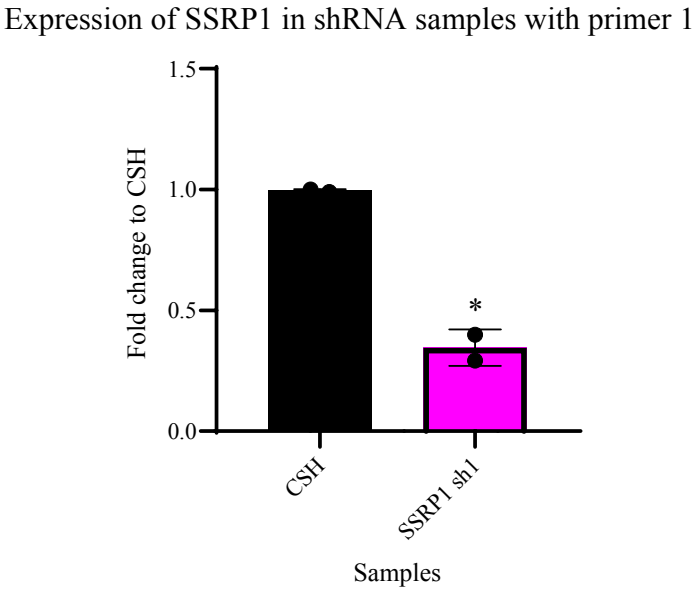
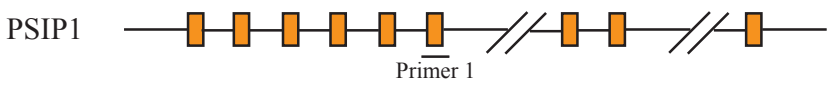


Figure 5.3: Validation of the expression of *SSRP1* gene knock-down with RT-qPCR.

(A) shows the positions of primer 1 on the *SSRP1* gene used in the qPCR step. The purple boxes represent the exons. (B) shows the graphs of *SSRP1* knock-down, plotted in relation to the fold change to CSH control. *SSRP1* expression was knocked-down in the SSRP1 shRNA 1 sample. (* <0.05, t-test, n=2 (Average value of two sets each performed in triplicate))

A



B

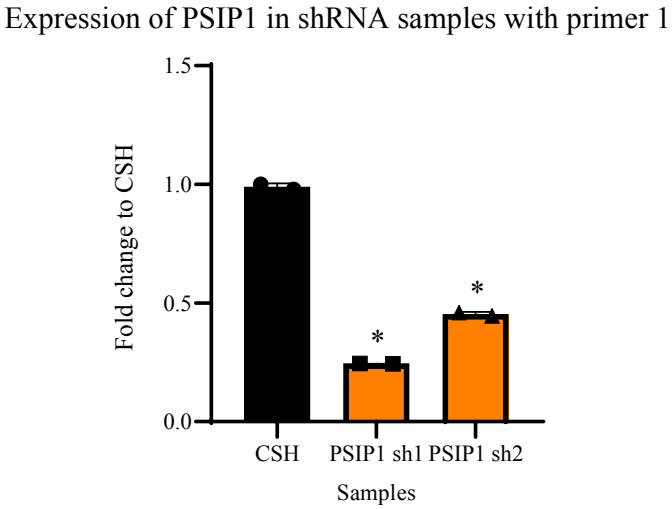


Figure 5.4: Validation of the expression of the *PSIP1* gene knock-down with RT-qPCR.

(A) shows the positions of primer 1 on *PSIP1* used in the qPCR step. The orange boxes represent the exons. (B) shows the graphs of *PSIP1* knock-down, plotted in relation to the fold change to CSH control. *PSIP1* expression was knocked-down in both PSIP1 shRNA 1 and shRNA 2 samples but to a greater extent in the shRNA 1 sample. (* <0.05, One-way Anova, Dunnett post-hoc test, n=2 (Average value of two sets each performed in triplicate)).

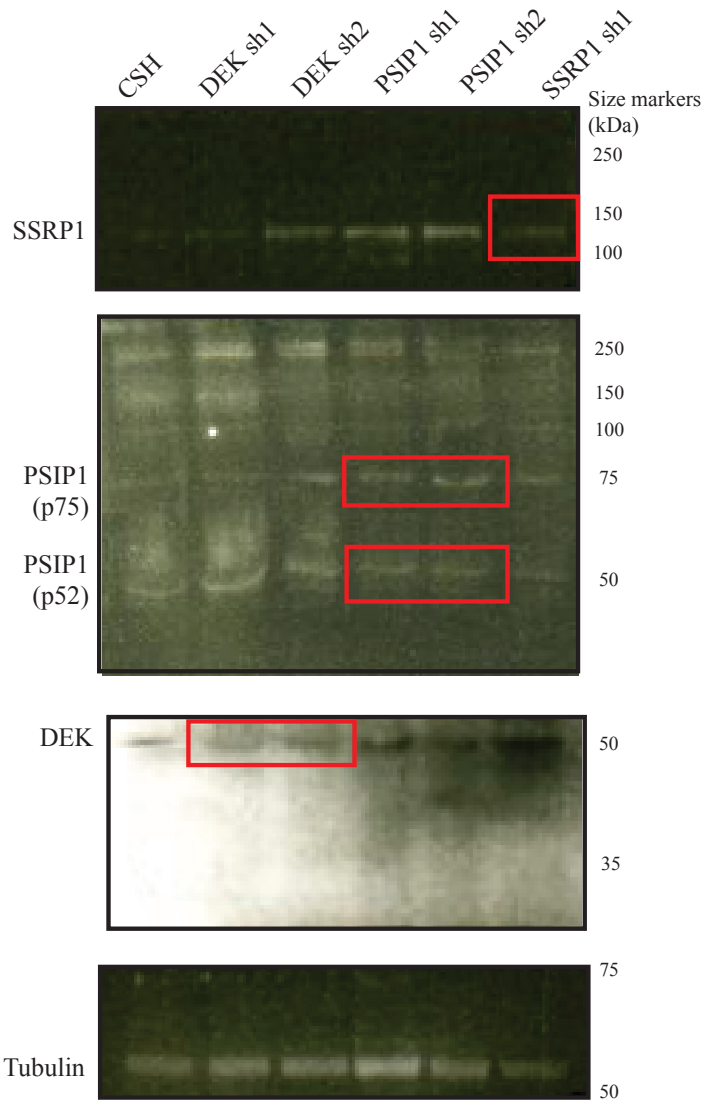


Figure 5.5: Protein expression of DEK, SSRP1 and PSIP1in shRNA samples. Proteins from shRNA samples were collected and western blots were performed. CSH was used as a control and tubulin as a loading control. The red boxes show the results for samples that correspond with the antibody used. No change in SSRP1 expression was observed in the SSRP1 shRNA-1 sample. The antibody used for detecting PSIP1 detected both isoforms. More PSIP1 p75 was expressed in the PSIP1 shRNA 1 sample than in the shRNA 2 sample, while expression of the p52 isoform was unchanged across the samples. For DEK, more expression was knocked-down by DEK shRNA 1 than by shRNA 2 or the CSH control.

5.2.4 Exploration of the effects of shRNA knock-down

5.2.4.1 The effects of shRNA knock-down on expression of *BCL11A*

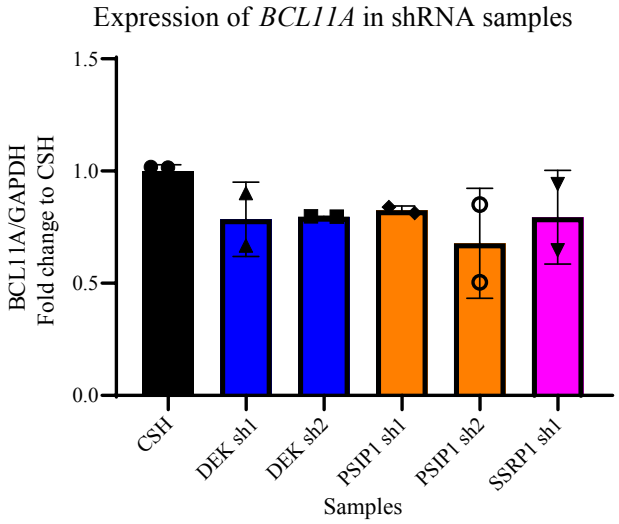
After the validation of the efficiencies of knock-down of the expressions of DEK, SSRP1 and PSIP1 with shRNAs, their effects on the expression of *BCL11A* were investigated. The qPCR result obtained is shown in Figure 5.6 A. No effect on expression of *BCL11A* was seen when DEK, SSRP1 or PSIP1 were knocked down individually. The y-axis of the graph was plotted in relation to the quantity of glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) and in relation to fold change of CSH. GADPH was used as an internal control as it is constitutively expressed in most cells and tissues. The knock-down of the expression of PSIP1 with both shRNA 1 and shRNA 2 resulted in the least expression of BCL11A among the shRNA samples. This is shown in the western blot displayed in Figure 5.6 B. Tubulin was used as loading control.

In addition to using the *GAPDH* gene as a control, the gene protein phosphatase 2 regulatory subunit B delta (*PPP2R2D)* was also used. *PPP2R2D* is responsible for the regulation of the cell cycle at the mitosis check-points [154] and was used as a housekeeping gene in the qPCR experiment. Housekeeping genes are important to maintain the basic cell functions for survival regardless of their roles [155]. Moreover, the primers of *PPP2R2D* were readily available from other work performed in our group. The qPCR result for this experiment is shown in Figure 5.7. No differences were observed across the shRNA samples tested.

5.2.4.2 The functional effects of shRNA knock-down

Apart from determining the effect of shRNA knock-down on the expression of *BCL11A*, the investigation was extended to study functional effects. Proliferation assay was performed first. This assay involved the seeding of 200 cells per sample and the calculation of the cell number every two days across 16 days. The result is shown in Figure 5.8A. The log of the cell number has been plotted against the number of days for easy visualisation and interpretation. No difference was seen in cell proliferation from the shRNA samples across the 16 days tested.

A



B

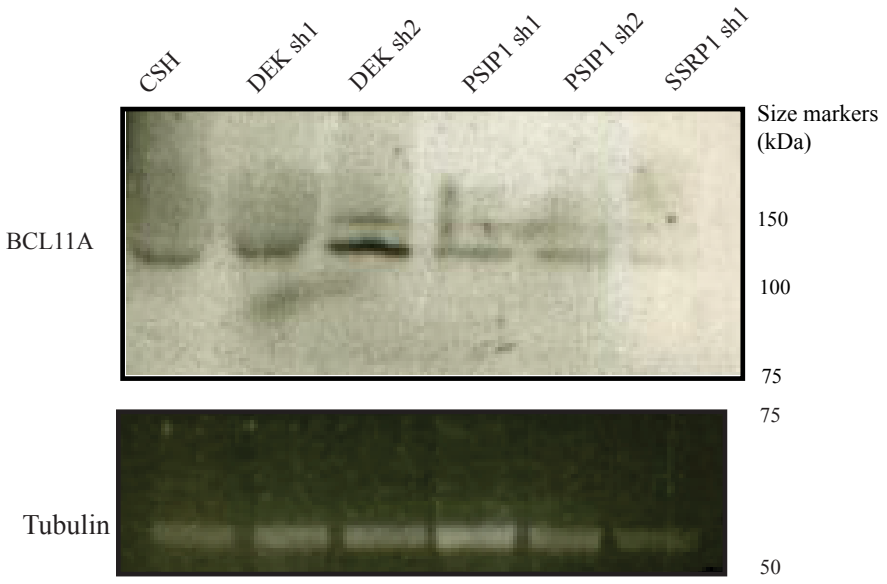


Figure 5.6: Evaluation of the effects of shRNA knock-down on expression of *BCL11A.*

The RT-qPCR graph in (A) reveals that there is no change in expression of *BCL11A* across the samples. (B) shows a western blot which indicates that knock-down of PSIP1 expression did affect protein expression of *BCL11A*. Tubulin was used as a loading control. n=2 (Average value of two sets each performed in triplicate)

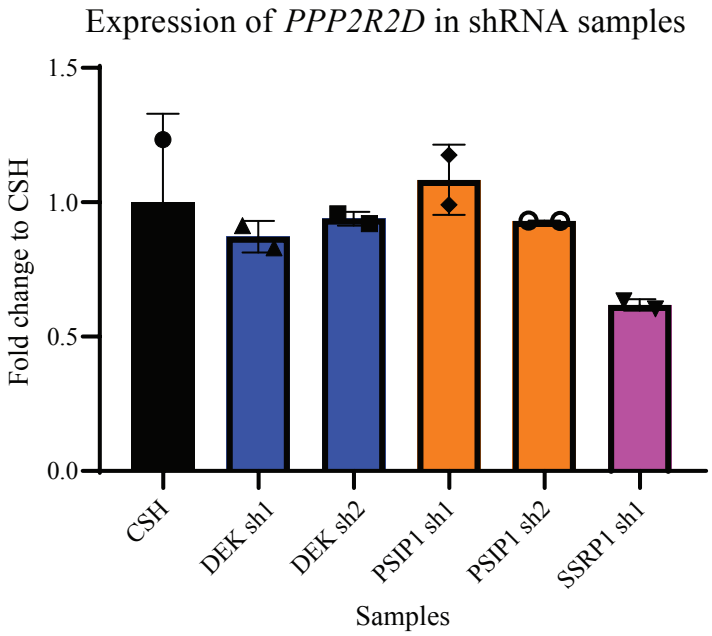


Figure 5.7: Evaluation of the effects of shRNA knock-down on global gene expressions.

Primers for *PPP2R2D* were used in the RT-qPCR. No effects on gene expression as a whole were observed after knock-down of DEK, SSRP1 or PSIP1 with shRNA. n=2 (Average value of two sets each performed in triplicate)

The knock-down of the expressions of DEK, SSRP1 and PSIP1 with shRNA on the effect of the potential to form colonies was also investigated. One hundred cells were seeded in the Matrigel and the numbers of colonies were counted after 14 days. As with the proliferation assay, it was observed in Figure 5.8 B that shRNA knock-down did not affect colony formation.

In summary, the knock-down of DEK, SSRP1 and PSIP1 did not affect the expression of *BCL11A* or influence the rate of cell proliferation or the potential to form colonies. One explanation for these results is that the three proteins do not work individually and that a combination of the three would be needed to regulate the expression of *BCL11A* or show functional effects. This hypothesis was tested by knocking-down the three proteins in combination.

5.2.5 Candidates work together to regulate expression of *BCL11A*

The schematic of the experimental plan to investigate the effect of the combination of the three proteins on expression of *BCL11A* is shown in Figure 5.9. The plan was similar to that used to knock-down the expressions when one shRNA was used per cell line. MDA-MB-231 cells were co-transfected with the shRNAs for DEK, SSRP1 and PSIP1. It was established in section 5.2.3 that the DEK shRNA 1 showed greater efficiency in knocking-down the expression of its target and PSIP1 shRNA 1 similarly was more efficient at knocking-down expression of PSIP1. The SSRP1 shRNA 1 plasmid was used in addition to the two mentioned. Transfected cells were selected using Feneticin and FACS to generate a stable cell line. The knock-down of the targets was then confirmed through the use of qPCR.

Firstly, the quantity of knock-down of expression of DEK was determined through the use of a pair of primers during the qPCR step. The positions of these primers pairs are shown in as previously in Figure 5.2. DEK expression was knocked-down in samples co-transfected with SSRP1 sh1 and PSIP1shRNA 1, but a greater degree was observed in the cells containing SSRP1 shRNA are shown in Figure 5.10. Cells co-transfected with DEK shRNA 1 and SSRP1 shRNA 1 or PSIP1 shRNA 1 were collected and the RNA was extracted. The qPCR result displayed in Figure 5.10 B

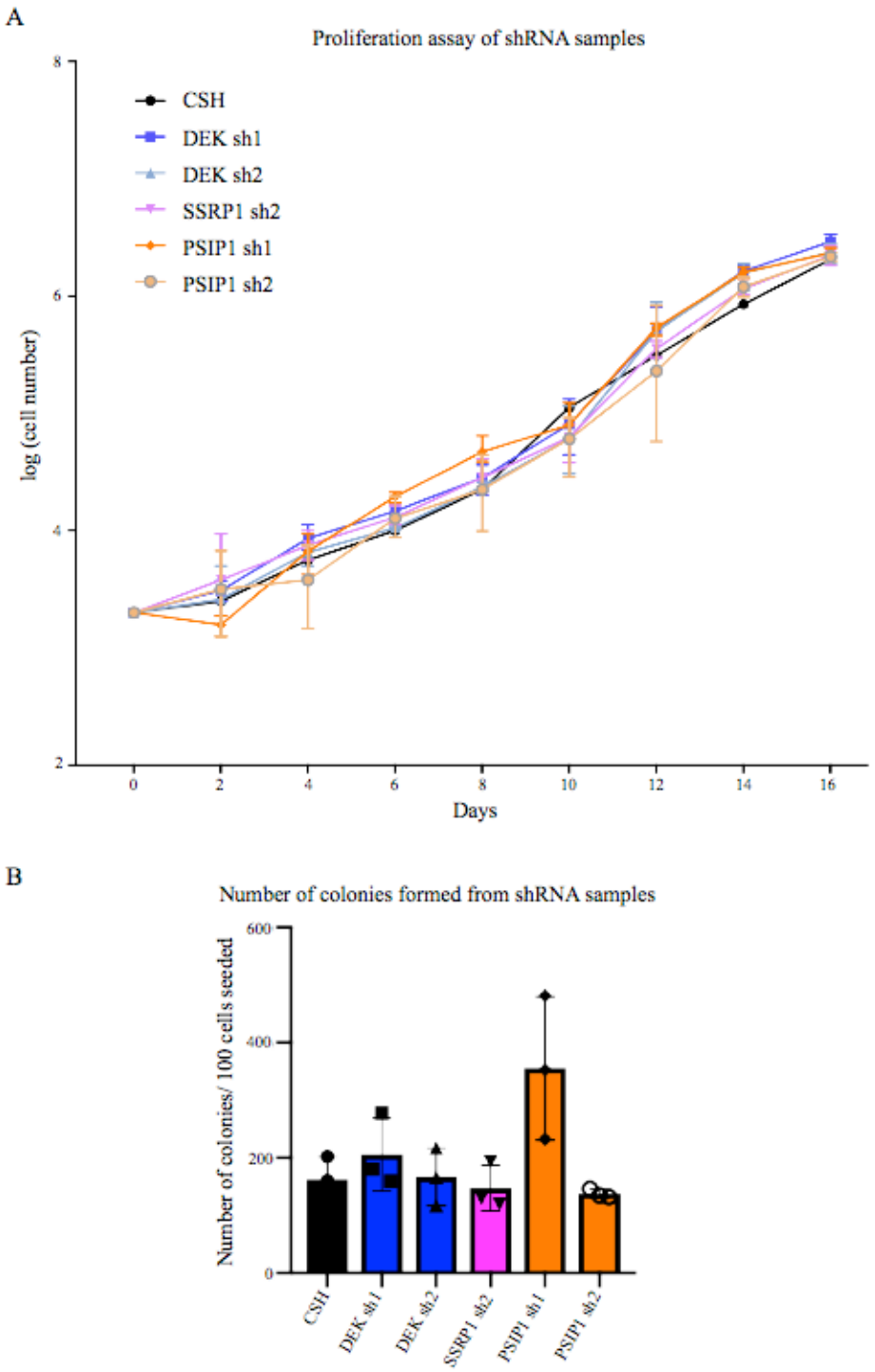


Figure 5.8: Functional studies of the knock-down of DEK, SSRP1 and PSIP1. The proliferation assay after 16 days is plotted as the log of cell number in (A). No difference in the rate of proliferation was observed among the shRNA samples, as shown in (B). n=3

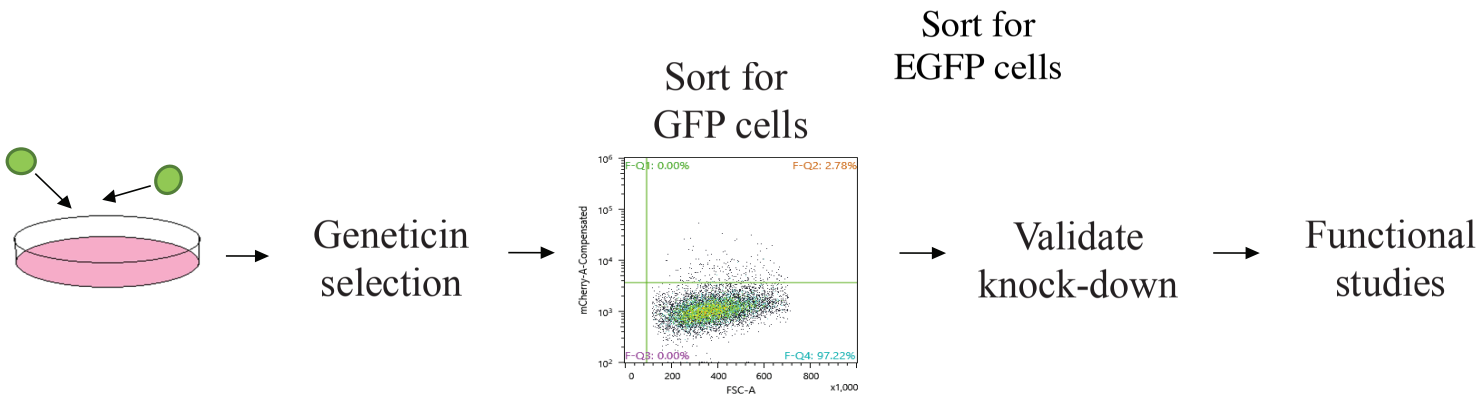


Figure 5.9: The schematic of the experimental plan for double shRNA knock-down.

MDA-MB-231 cells were first co-transfected with plasmids that expressed shRNA for the three different targets. Cells were then selected by geneticin and sorted for EGFP-positive cells to generate stable cell lines. Cells were then collected to validate knock-down and perform functional studies.

show that expression of DEK has been successfully knocked-down in both samples. Secondly, we determined the knock-down of SSRP1 in cells co-transfected with SSRP1 shRNA 1 and DEK shRNA 1 or PSIP1 shRNA 1. The cells were collected and RNA extracted. The positions of the primers used in the qPCR experiment are shown in Figure 5.11 A and the results in 5.11 B. The expression of SSRP1 was found to be knocked down by the SSRP1 shRNA 1 plasmid in both cell lines.

Lastly, cells co-transfected with PSIP1 shRNA 1 and DEK shRNA 1 or SSRP1 shRNA 1 were collected to determine the expression of PSIP1. Again, the positions of the primers used in the qPCR step is shown in Figure 5.12 A. Cells co-transfected with PSIP1 shRNA 1 and DEK shRNA 1 or SSRP1 shRNA 1 plasmids were found to contain knocked-down PSIP1 expression in Figure 5.12 B. A greater degree of knock-down was observed in samples co-transfected with PSIP1 shRNA 1 and DEK shRNA 1.

Protein lysate in the cells co-transfected with two shRNA plasmids were collected and western blot was performed. The result in Figure 5.13 showed that the knock-down of DEK in samples co-transfected with DEK shRNA 1 and SSRP1 shRNA 1 or PSIP1 shRNA 1 was reflected at the protein level. Similarly, SSRP1 knock-down was seen in cells co-transfected with SSRP1 shRNA 1 and DEK shRNA 1 or PSIP1 1 and DEK shRNA 1. For PSIP1, no change was observed in the levels of the PSIP1 p75 isoform across the samples, but the p52 isoform was only observed in samples co-transfected with PSIP1 shRNA 1 and SSRP1 shRNA 1.

5.2.5.1 Double shRNA knock-down affects expression of *BCL11A*

The effect on the expression of *BCL11A* of the co-transfection of three cell lines with two shRNA plasmids was compared with CSH and the results are shown in Figure 5.14 A. Knock-down of multiple genes resulted in a greater than 60 per cent reduction in the expression of *BCL11A* in all the cell lines tested. Among them, double knock-down with SSRP1 shRNA 1 and PSIP1 shRNA 1 plasmids showed the greatest reduction in expression of *BCL11A.* To determine whether this knock-down would show a global effect, we used *PPP2R2D* as a control. The results in

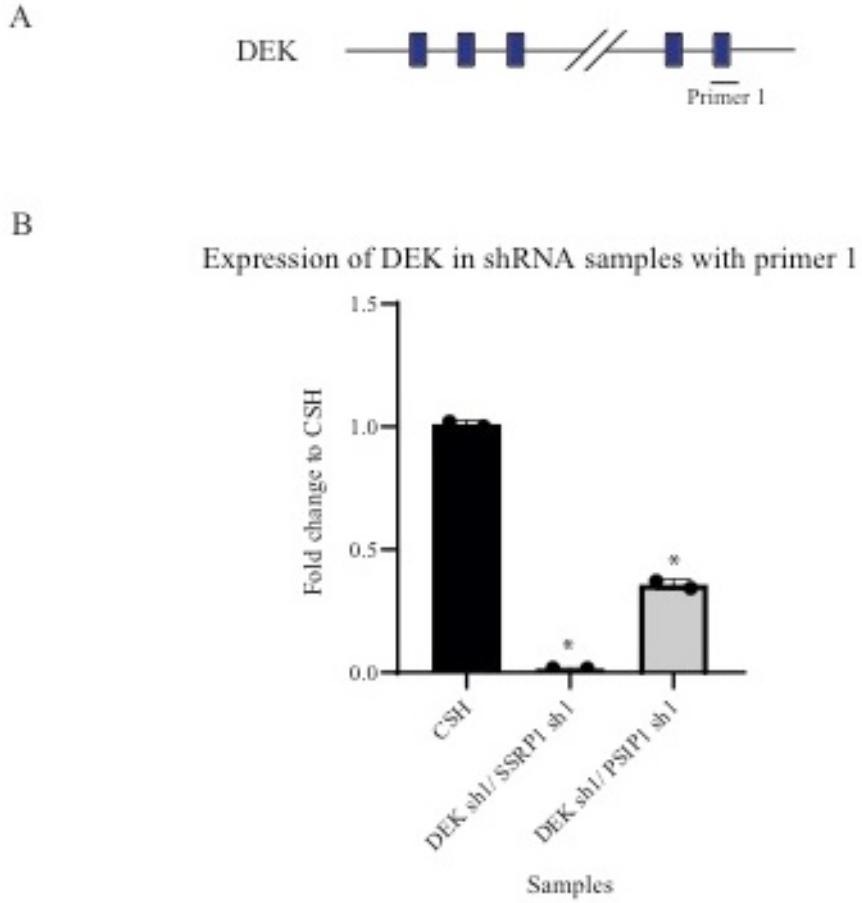


Figure 5.10: Validation through RT-qPCR of the knock-down of expression of *DEK* in double knock-down samples. .

(A) is a graphical representation of the positions of primers on *DEK* used in qPCR.. The blue boxes represent the exons. *DEK* expression was knocked-down in both samples as shown in (B) and (C) but to a greater extent in the DEK shRNA 1/SSRP1 shRNA 1 sample. (*<0.05, One way Anova, Dunnett post-hoc test, n=2 (Average value of two sets each performed in triplicate))

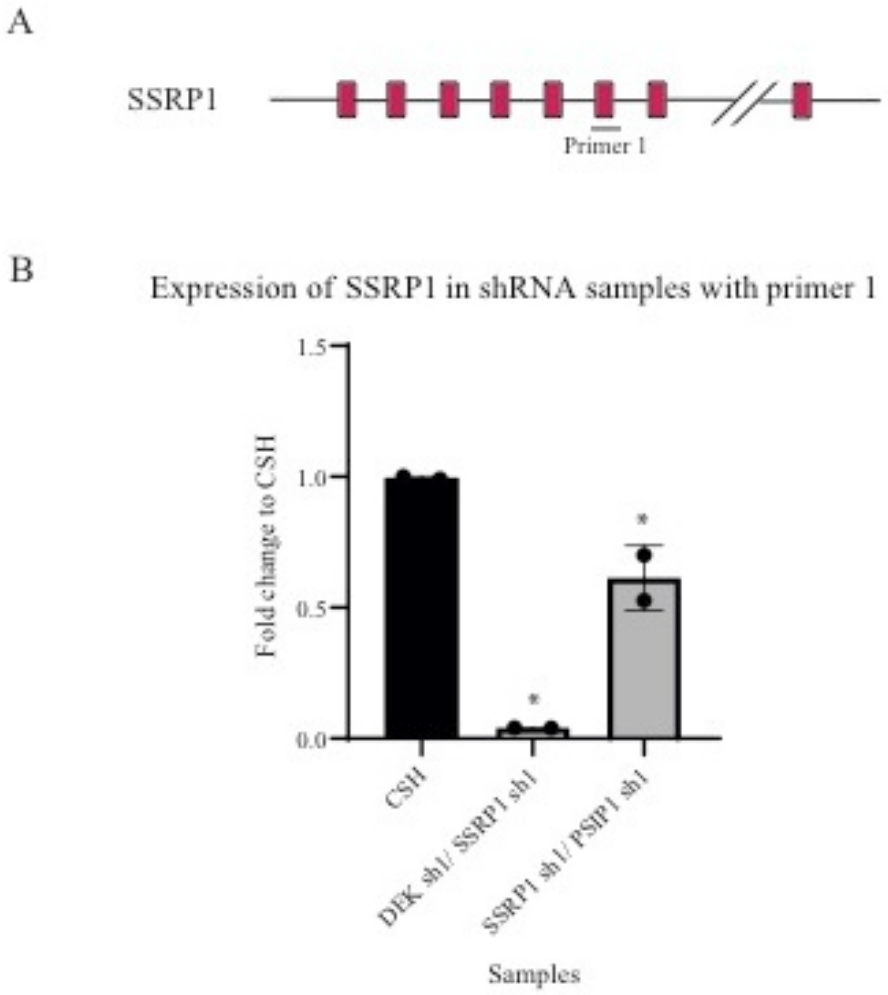


Figure 5.11: Validation through RT-qPCR of the knock-down of expression of *SSRP1* in double knock-down samples.

(A) is a graphical representation of the positions of primers on *SSRP1* used in qPCR. The purple boxes represent the exons. *SSRP1* expression was knocked-down in both samples shown in (B) and (C) but to a greater extent in cells co-transfected with DEK shRNA 1/SSRP1 shRNA 1. (*<0.05, One way Anova, Dunnett post-hoc test, n=2 (Average value of two sets each performed in triplicate))

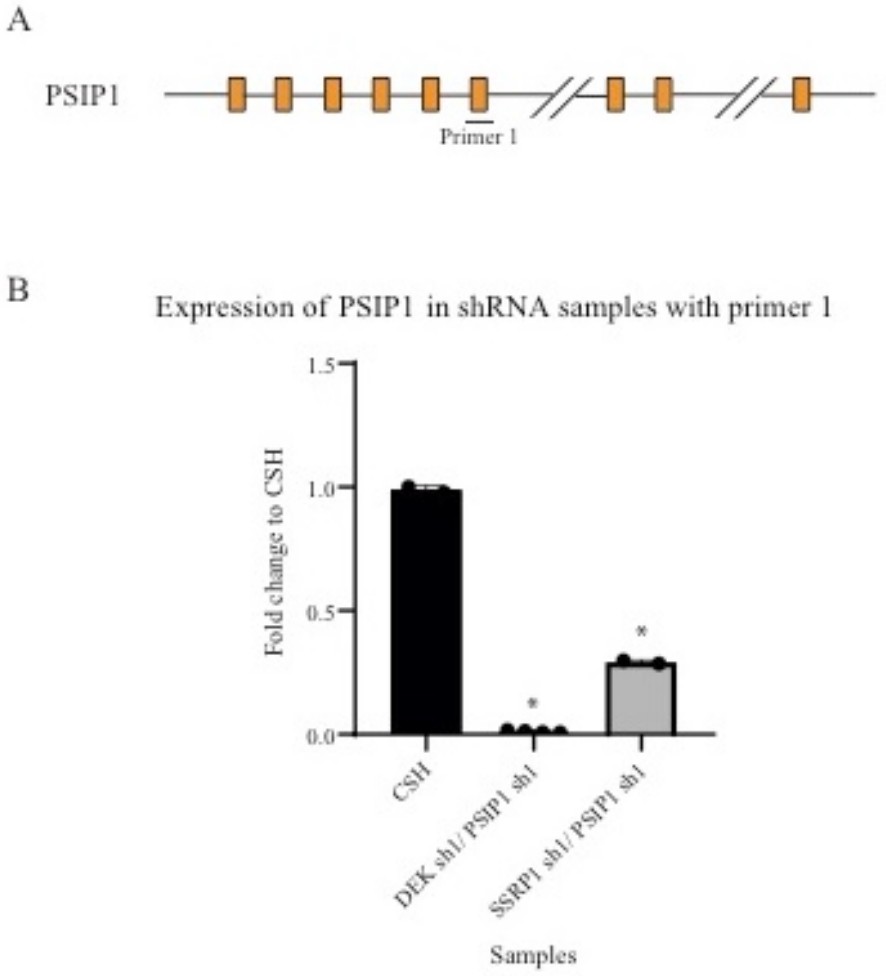


Figure 5.12: Validation through RT-qPCR of the knock-down of expression of *PSIP1* in double knock-down samples.

(A) is a graphical representation of the positions of primers on *PSIP1* used in qPCR. The orange boxes represent the exons. *PSIP1* expression was knocked-down in the DEK shRNA 1/PSIP1 shRNA 1 sample using both primers, but only with primer 2 in cells co-transfected with SSRP1 shRNA 1/PSIP1 shRNA 1. (*<0.05, One-way Anova, Dunnett post-hoc test, n=2 (Average value of two sets each performed in triplicate))

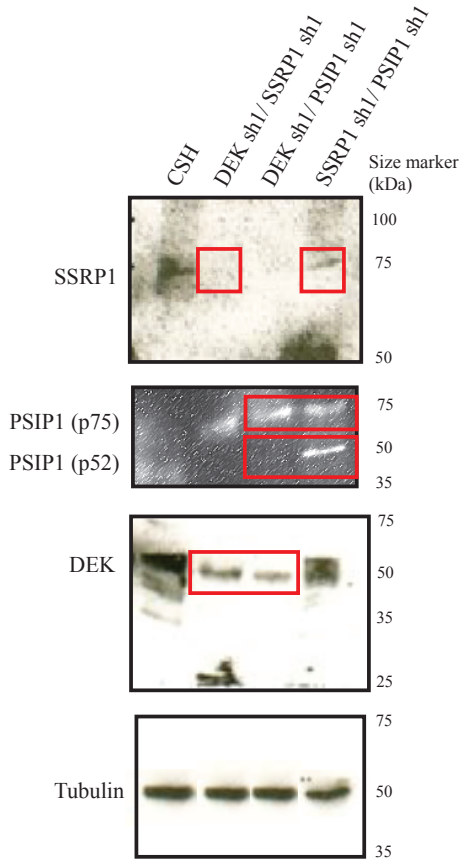


Figure 5.13: Protein expression for DEK, SSRP1 and PSIP1 in double knock-down samples.

Lysates of proteins produced in double knock-down samples were collected and western blot was performed. CSH was used as a control and tubulin as a loading control. The red boxes show samples that correspond with the antibody used. SSRP1 expression was knocked-down in cells co-transfected with DEK shRNA 1 or PSIP1 shRNA 1. A greater degree of SSRP1 knock-down was observed in cells with DEK shRNA 1 and SSRP1 shRNA 1. No change was observed in the PSIP1 p75 isoforms among the double knock-down samples, but the PSIP1 p52 isoform was only seen in the sample co-transfected with SSRP1 shRNA 1 and PSIP1 shRNA 1. Knock-down of DEK was performed in samples co-transfected with DEK shRNA 1 and SSRP1 shRNA 1 or PSIP1 shRNA 1.

Figure 5.14 B show that double knock-down with DEK shRNA 1 and SSRP1 shRNA 1 or PSIP1 shRNA 1 had a global effect on gene expression, while this effect was absent when SSRP1 and PSIP1 were knocked-down together.

Protein lysates were collected from these cell lines to investigate whether these knock-down effects were translated to the protein level. The western blot in Figure 5.15 shows that expression of BCL11A was not seen at the protein level in any of the double knock-down samples.

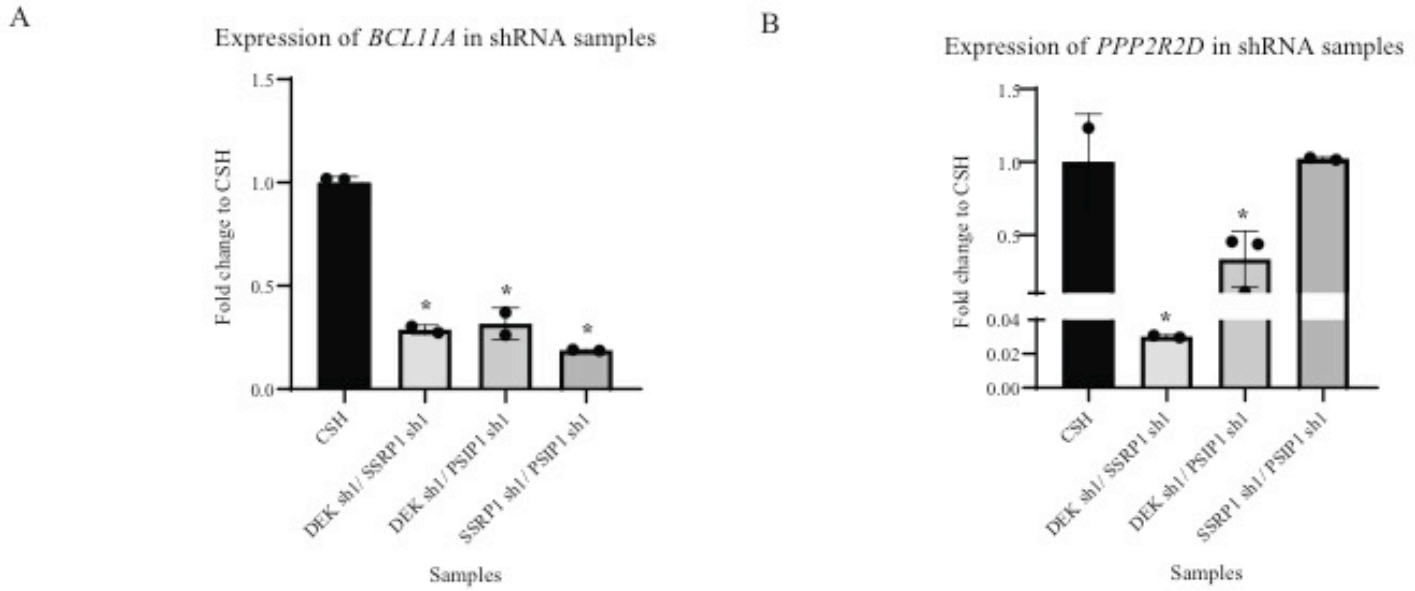


Figure 5.14: Evaluation of the effects of double shRNA knock-down on expression of *BCL11A* and *PPP2R2D*.

(A) shows the RT-qPCR results of the expression of *BCL11A* in double shRNA knock-down samples. All three double shRNA samples showed a greater than 60 per cent knock-down of *BCL11A* expression and this was also reflected at the protein level. (B) shows the RT-qPCR results of the expression of *PPP2R2D* in double shRNA knock-down samples. A global effect was seen in samples co-transfected with DEK shRNA /SSRP1 shRNA 1 and DEK shRNA 1/PSIP1 shRNA 1. * <0.05, One-way Anova, Dunnett post-hoc test, n=2 (Average value of two sets each performed in triplicate)

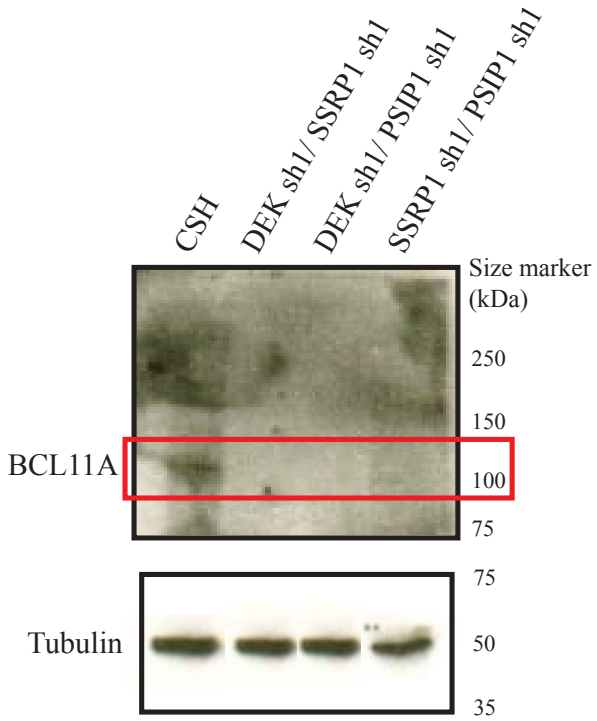


Figure 5.15: Protein expression for BCL11A in double knock-down samples.

Lysates of proteins produced in double knock-down samples were collected and western blot performed. CSH was used as a control and tubulin as a loading control. BCL11A expression was knocked-down in cells co-transfected with DEK shRNA 1/ SSRP1 shRNA 1, /PSIP1 shRNA 1 and SSRP1 shRNA 1/ PSIP1 shRNA 1.

5.3 Discussion

5.3.1 Evaluation of the experimental design

To understand the ways in which DEK, SSRP1 and PSIP1 could regulate the expression of *BCL11A*, we obtained shRNA primers from the MISSION® Sigma database and determined their specificities for our genes of interest by blasting them against the data contained in the UCSC genome browser. The results produced confidence that these primers targeted DEK, SSRP1 or PSIP1. They were then cloned into the shRNA-expressing plasmid together with primers that contained complementary sequences. Two shRNAs for DEK and PSIP1 each were cloned successfully while one was cloned for SSRP1. Next, these plasmids were transfected into MDA-MB-231 cells using the PB transposon system, as described in Chapter 3. The selection markers, EGFP and the geneticin-resistance gene, in the shRNA-expression plasmid enabled selection of transfected cells.

After generating stable cell lines which constitutively expressed the shRNA for the targets, cells were collected, RNA extracted and qPCR performed to determine whether the designated genes had been knocked down. The strategy employed to determine the degree of knock-down was use of one specific primer per gene during the qPCR step. The results shown in Figures 5.2- 5.4 demonstrate that the expressions of DEK, SSRP1 and PSIP1 have been knocked down by their corresponding shRNAs. Moreover, it was observed from the transfected cell lines that the DEK shRNA 1, SSRP1 shRNA 1 and PSIP1 shRNA 1 plasmids were more effective at knocking-down their target genes in comparison with the other designed shRNAs. The knock-down of DEK and SSRP1 in cells were both reflected at the protein level. This could be due to the positions of the target regions on DEK and SSRP1 and their potential to influence the stability of the protein or protein folding.

5.3.2 Functional effects of shRNAs

Following the validation of the methods of knock-down employed for expression of mRNA, functional assays were set up to study the effects. The choice of function studies was determined by the known functions of DEK, SSRP1 and PSIP1 in breast

cancer. Firstly, the effects on expression of *BCL11A* were investigated, as the aim of this project was to identify potential regulators that might influence the expression of *BCL11A*. Moreover, DEK and PSIP1 proteins are known to regulate transcription [129] [135] [156] [157], while SSRP1 is known to facilitate transcription by disassociating the histone proteins and is known to act as a subunit for the FACT complex [131]. This complex is also important for regulating transcription.

The qPCR results displayed in Figures 5.6 show that knocking down the three proteins did not affect the expression of *BCL11A* at the RNA level. These results also translate at the protein level, except in the case of cells transfected with PSIP1 shRNA. The knock-down of the expression of PSIP1 has been shown to have resulted in the reduction of protein expression by *BCL11A*. This suggests that PSIP1 might be important in the expression of *BCL11A*.

GAPDH was used as an internal control for experiments in which we have determined the effect of knocking down the three proteins on expression of *BCL11A*. In addition, *PPP2R2D* was chosen as a second control. *PPP2R2D* was used due to its function as a housekeeping gene whose expression did not vary in tissues and cells and which was responsible for the regulation of the cell cycle at the mitosis check-points [154]. It can be seen from the results shown in Figure 5.7 that knocking down the expressions of DEK, SSRP1 and PSIP1 did not affect the expression of *PPP2R2D*. This suggests that these proteins may regulate gene expression selectively and do not affect other genes on a global scale.

Proliferation and colony assays were also performed. It is known that DEK, SSRP1 and PSIP1 perform functions in breast cancer, including the formation of mammosphere [138] and involvement in transformation [142] [143] and migration [138] [139] [144] . Other functions were summarised in the discussion section of Chapter 4. This work has shown that knocking down the three genes did not affect the rate of cell proliferation or the cells' ability to form colonies. These results suggest that these proteins may show selectivity in their functions in different breast cancer subtypes. Another potential explanation is that DEK, SSRP1 and PSIP1 do

not act alone. To investigate this hypothesis, we co-transfected various combinations of shRNAs to target two different mRNA transcripts at the same time.

5.3.3 The effect of double shRNA knock-down

As in single shRNA transfection experiments, cells with two shRNAs were co-transfected, then selected through geneticin and FACS sorting. Cells from the stable cell lines were collected, RNA extracted and qPCR performed to determine the degree of gene knock-down.

We will first discuss the results of knocking-down in cells with two shRNAs. Firstly, cells co-transfected with DEK shRNA. DEK was knocked-down at both RNA and protein levels in cells that had been co-transfected with SSRP1 sh1 or PSIP1 sh1 plasmids. Among these samples, cells co-transfected with SSRP1 shRNA 1 showed the greatest level of knock-down in comparison with cells transfected with DEK shRNA alone. Secondly, similar result was observed that a greater level of SSRP1 was knocked-down in cells co-transfected SSRP1 shRNA 1 with DEK shRNA 1 than PSIP1 shRNA 1. Lastly, the same was seen in the samples co-transfected PSIP1 shRNA 1 with DEK shRNA 1. We have observed the same trend in these results, which is a greater level of knock-down when cells were co-transfected with DEK shRNA 1.

After the evaluation of the degree of knock-down in the expression of the targets by their corresponding shRNAs, we have also investigated the impact of shRNAs on expression of *BCL11A*. It was surprising to find that through the combination of knock-down of two genes at the same time, a great reduction in the expression of *BCL11A* was observed at both RNA and protein levels. All three double knock-down cell lines showed a reduction of over 60 percent in the expression of *BCL11A* in comparison with the CSH sample. Among them, the greatest effect on *BCL11A* expression was observed in the experiments in which *SSRP1* and *PSIP1* were knocked-down. In addition, we have shown through the *PPP2R2D* qPCR results that a global effect is not observed in this cell line. These results suggest that SSRP1 and PSIP1 may work together and are more selective in the regulation of expression

of *BCL11A* than are other genes, in comparison with other double knock-down combinations. In addition, a common trend was also noticed in the *PPP2R2D* qPCR results– cells co-transfected with DEK shRNA 1 will result in a global effect on gene expressions. These results suggest that DEK might influence global gene expression including SSRP1 and PSIP1.

5.3.4 Proposed mechanism for regulation of expression of *BCL11A* by DEK, SSRP1 and PSIP1

A combination of these results with known literature leads to a proposed mechanism of ways in which DEK, SSRP1 and PSIP1 may regulate the expression of *BCL11A*. In the single shRNA knock-down experiments, the knock-down of *DEK* did not have any effect on expression of *BCL11A*. However, knock-down in combination with *SSRP1* and *PSIP1* significantly reduced expression of *BCL11A* at both gene and protein levels. This can be explained by consideration of the known functions of SSRP1 and PSIP1 as transcription co-activators. Therefore, they might recruit DEK to regulate transcription. In addition, the qPCR result in *PPP2R2D*, shown in Figure 5.14, suggests that knock-down of *DEK* affects general gene expressions rather than those *BCL11A* specifically. These results suggest that DEK is a general regulator for transcription, and this matches its known functions.

Secondly, the knock-down of *SSRP1* alone had no effect on *BCL11A*'s gene or protein expressions. However, a combination of shRNA knock-down of *SSRP1* with knock-down of *DEK* or *PSIP1* significantly reduced both gene and protein expression of *BCL11A*. These results suggest that SSRP1 requires the presence of additional proteins to regulate transcription of *BCL11A* and may act as a transcriptional co-activator. The proposed mechanism is shown in Figure 5.16 A.

Lastly, single shRNA knock-down of *PSIP1* had no effect on gene expression by *BCL11A* but produced a partial reduction at the protein level. On the other hand, combination knock-down of *PSIP1* with *DEK* or *SSRP1* led to a significant reduction in both *BCL11A*'s gene and protein expressions. These results led to the proposal, shown in Figure 5.16 B, that PSIP1 may act in a similar way to SSRP1 as

a transcriptional co-activator but through the recruitment of different transcription factors. PSIP1 also regulates protein expression of *BCL11A*.

In addition, the knock-down of *PSIP1* and *SSRP1* together had no effect on global gene expression, in contrast with the combination knock-down of *DEK* with *SSRP1* or *PSIP1*. It is also worth noting that the knock-down of *SSRP1* alone affected general gene expression, i.e. these cells do not express *SSRP1* but still express *PSIP1*. These results indicate that PSIP1 may negatively regulate global gene expressions, and its presence in a mixture may negatively regulate the recruitment of SSRP1 and other associated factors.

5.4 Conclusion

To conclude, this chapter details the establishment of cell lines that contain one or two types of shRNA molecules that target DEK, SSRP1 or PSIP1. The knock-down of these proteins individually does not affect the expression of *BCL11A,* nor does it affect the functional assays. However, we have shown that combinations of either two of these proteins are needed to regulate gene expression. Among them, SSRP1 and PSIP1 might work together to play a bigger role as regulators of *BCL11A,* in comparison with mixtures of DEK with SSRP1 or with PSIP1. These results imply that DEK, SSRP1 and PSIP1 could be potential targets in the development of novel therapeutic targets against *BCL11A* in TNBC.

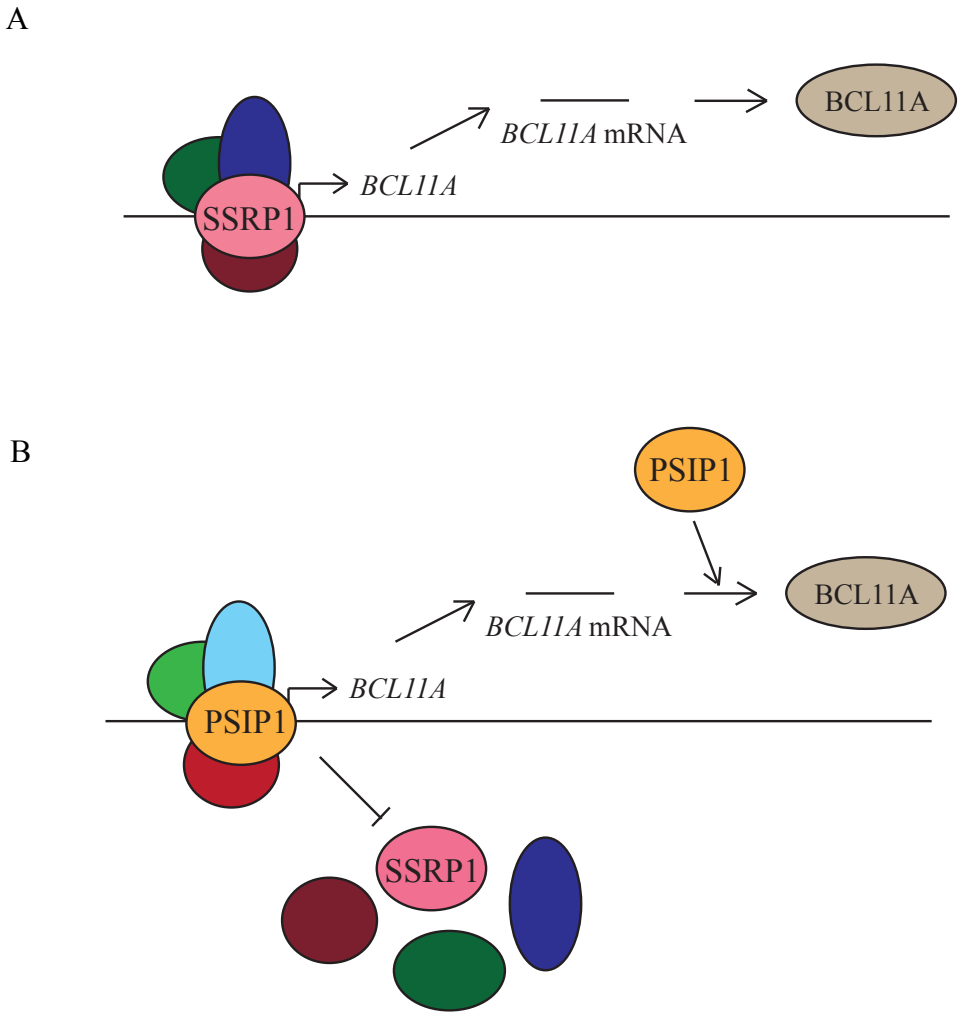


Figure 5.16: Proposed mechanisms of regulation of *BCL11A*'s expression.
(A) shows the proposed mechanism of action by SSRP1, which acts as a co-transcriptional factor to recruit other transcription factors to regulate the expression of *BCL11A*. (B) shows the speculated mechanism of action for PSIP1. It indicates a number of different roles. Firstly, it acts as a transcriptional co-activator that recruits different transcription factors to those recruited by SSRP1. Secondly, it takes part in the regulation of protein expression by *BCL11A*. Finally, it negatively regulates global gene expressions and hence negatively regulates SSRP1's expression.

CHAPTER 6

Discussion, future directions, and conclusion

6.1 Discussion

6.1.1 Rationale behind the research aim and Achievements

The main aim of this study is to investigate the transcriptional regulation of *BCL11A* in Triple-negative breast cancer. The chief reason for conducting this project comes from the clinical challenges in treating this breast cancer subtype. Clinical characteristics of TNBC include high mortality rates and recurrence within 5 years [42] in comparison to other subtypes. The lack of hormonal and HER-2 growth factor receptors means that conventional therapies are unsuitable for treating TNBC patients and chemotherapy is the only treatment available. These call for the identification of new molecular markers and drug treatments for TNBC patients. Advancement on TNBC's research has led to the identification of *BCL11A* as an oncogene and our group has demonstrated its role in initiating and maintaining tumourigenesis [74]. *BCL11A* is known to be involved in many biological functions. In the immune system, it silences γ-globin chains and participates in the complex that governs globin chain switching during development [62] [64] [65] [66] and regulates the development of T- and B-cells [63]. It is also responsible for regulating axon branching and dendrite growth [119] in the neural system. Its importance in these biological systems means that developing an inhibitor against *BCL11A* directly was unfavourable and the identification of potential regulatory factors in this project would shed light on drug development for TNBC patients against this oncogene.

I utilise the CRISPR-Cas9 system coupled with RIME proteomics to investigate potential factors regulating *BCL11A*'s expression in this project. The importance of the sites I chose to study was evaluated in Chapter 3; namely, I have identified DEK, SSRP1, and PSIP1 as potential putative regulators of *BCL11A* from RIME proteomics in Chapter 4, and I then demonstrated that SSRP1 and PSIP1 work in concert to regulate *BCL11A* in TNBC using a RNA interfering technique in Chapter 5.

The specific use of the CRISPR-Cas9 system in my project is novel in two key aspects, compared to conventional implementation. Firstly, dCas9 is used instead of normal Cas9, and repurposed to enable targeting of specific sites of *BCL11A* while avoiding DNA double-strand breaks that would otherwise be introduced. Secondly, rather than using dCas9 as an enzyme, I use dCas9 as a protein tag that is pulled down by specific antibodies. This allows us to identify and study DEK, SSRP1, and PSIP1 as regulators for *BCL11A*'s expression in TNBC at our sites of interest.

6.1.2 Evaluating the CRISPR-Cas9 system

The CRISPR-Cas9 system is used to study transcriptional regulations at specific *BCL11A*'s sites. This system was originally identified as a defense mechanism against foreign DNA by bacteria and archea using a RNA-guided DNA recognition system [89] [90]. This was achieved by transcribing foreign DNA from the host's genome incorporated during the last encounter. This transcript, called crRNA, recognises complementary sequences from invaders upon a second encounter, guides Cas9 to the site and induces DNA double-strand breaks [95]. Section 1.6.1.1 described the detailed mechanism. Since its discovery, the CRISPR-Cas9 system has been adapted for genetic engineering and genome editing [103] [104] [105] [106] [109] [110].

Taking advantage of the strengths of this system, I designed gRNA, which acts in the same capacity as crRNA in bacterial defence systems, to target specific *BCL11A* sites (Sites A and B) in my project. These target sites are selected from the UCSC genome browser as active transcription sites through display of the H3K27Ac histone marker and hypersensitivity to DNaseI. I validated this using a normal Cas9 to delete these target sites. Deletion of entire Sites A and B was initially attempted but cells did not survive during selection. Therefore, I switched to deleting smaller fragments within these sites. This could have allowed me to identify smaller sites within Sites A and B that regulate the expression of *BCL11A*. Results in Figure 3.2 suggest that Sites A and B, while important for *BCL11A*'s expression, also affects downstream pathways and deleting them entirely was thus detrimental to cell

survival. This could be further investigated with functional assays, for example a survival assay.

Apart from using normal Cas9, I also employ dCas9 to target *BCL11A* sites instead of inducing DNA double-strand breaks and use this protein as a tag for pulling down itself along with proteins in close proximity for the RIME experiment. This dCas9 is expressed with an EGFP fluorescence signal through a self-cleavable 2A protein. In addition, I use the Tet-On system to induce these proteins. Data from Figure 3.5 show successful induction of dCas9 using Tetracycline derivative, Doxycycline, and illustrate how the EGFP level correlates with expression of the dCas9 protein. One disadvantage of the CRISPR-Cas9 system is the possibility of CRISPR interference— the blockage of genetic regions and accessibility of proteins by Cas9 [110]. This is particularly relevant to my project as the chief aim is precisely to study transcription regulation. To evaluate this effect, I perform a timed induction experiment by collecting RNA at various time-points within 24 and 48 hours of Doxycycline induction. Results in Figure 3.5 show that dCas9 induction does not affect the expression of *BCL11A* regardless of the time of induction. I also evaluate the specificity of dCas9 by performing ChIP experiments using specific Anti-V5 antibodies against the multi-tagged dCas9. Results in Figures 3.6 and 3.7 show binding specificities at both Sites A and B. In the future, ChIP-sequencing experiments should be considered to evaluate the global binding of dCas9 on the human genome. This data will provide valuable insights on the binding and off-target effects on a global scale.

In order to identify transcriptional regulators which are upregulated or have specific expressions in TNBC, I attempted to perform this CRISPR-Cas9 system on Luminal cell lines MCF7 and T47D. Unfortunately, there was little success with transfection and establishment of stable cell lines proved to be challenging.

6.1.3 Evaluating the experimental design and proteomic analysis from the RIME experiment

I discussed in Chapter 4 that DEK, SSRP1, and PSIP1 are identified from the RIME analysis. RIME is a technique developed to study unknown proteins, in the context of this project those which regulate transcription of *BCL11A*. It couples ChIP which pulls down dCas9 as well as proteins sitting in close proximity, with mass spectrometry which identifies these proteins. This technique also allows the study of endogenous proteins in their native state through the fixation of stable cell lines established in Chapter 3 with formaldehyde, conserving even transient interactions [83]. The nuclei are then isolated in an attempt to identify proteins with functional roles in regulating transcription which takes place inside.

326 and 491 proteins were identified to sit close to dCas9 at Sites A and B, respectively. In order to make the list more manageable, they are filtered by multiple criteria detailed as follows. Firstly, proteins that overlap with the controls are eliminated and those between or within Sites A and B samples are processed. Next, they are selected using parameters related to the mass spectrometry step as a form of quality control. Proteins with a unique peptide of one are selected and this gives confidence that the 'correct' protein is identified. The remaining list of proteins is sorted from high to low according to a cumulative score, based on the aggregation of scores given to each protein from the RIME database; the higher the score, the more abundant it was in the sample. Then, to identify proteins that may regulate the expression of *BCL11A*, I select proteins that were expressed in the nucleus to proceed to the next criteria. These selected proteins are compared against an established database called Crapome, allowing me to remove common protein contaminants from proteomic studies. They are further compared against the cancer database, cBioPortal, in terms of their mRNA expressions against the PAM_{50} and correlation to *BCL11A*. PAM_{50} is an assay for classifying intrinsic breast cancer subtypes by quantifying the co-expression of 50 genes [27]. Lastly, only proteins with known roles in breast cancer are selected for further studies. These are DEK, SSRP1, and PSIP1.

To evaluate if these three factors are meaningful pull-downs, I first validate the cBioPortal data in which they are found to be upregulated in Basal-like breast cancer; 80% of this subtype is equivalent to TNBC. Proteins are collected from a panel of TNBC and Luminal cell lines and western blots are performed. Results reported in Figure 4.4 agree with the cBioportal database. However, their expression in Luminal as well as TNBC cell lines suggest that these proteins have general roles. Other than their subtype expressions, the binding of DEK, SSRP1 and PSIP1 at the target sites and regions nearby are also validated in MDA-MB-231 and HCC1569 TNBC cell lines. Similar to the ChIP-qPCR experiment described in Chapter 3, specific antibodies against factors are used to pull-down the DNA-bound proteins and these DNA regions are in turn amplified by specific primers. Figures 4.5- 4.8 confirm protein binding at Sites A and B and nearby regions. These results also verify that this RIME proteomics technique is applicable to the identification of unknown proteins.

In a database called STRING, DEK is said to interact with SSRP1 and PSIP. I attempt to validate this in MDA-MB-231 cells by Co-Immunoprecipitation experiments using DEK as the central protein and for pulling down the other two factors. However, this proves to be challenging as the molecular weight of DEK is very close to the heavy chain of the Protein-G beads used. To resolve this, I switch to using Protein-A, which has a smaller molecular weight than DEK, as resin. Another merit of using Protein-A comes from its higher sensitivity for human Rabbit antibodies. From the Co-IP, I find that SSRP1 and PSIP1 are absent in the pulled-down samples. One possible explanation for this is that the MDA-MB-231 wild type cells had little SSRP1 and PSIP1 expression and hence the majority of binding sites on the DEK antibodies is not occupied. To remedy this, Co-IP is performed in HCC1569 wild type cells which has been shown previously to have high expression of the three factors. Regrettably, I was unable to finish optimisations due to time constraints and was therefore unable to conclude if these three factors form a complex as suggested by the STRING database.

6.1.4 Implications of RIME pull-down

Chapter 5 discussed the functional effects of DEK, SSRP1, and PSIP1, as well as their impacts on *BCL11A*'s expression in TNBC. These factors have known roles in cells and breast cancer, as reviewed in Table 4.11. I will briefly summarise and discuss their implications on my results here.

DEK has been identified to organise chromatin which helps regulate transcription, for mRNA splicing, and for DNA replication and repair. In terms of expression, DEK protein is dependent on tRNA post-transcriptional modifying enzymes which in turn promote IRES-dependent translation for proteins such as LEF1 [152]. This study by Delaunay et al showed that such protein translation stimulates breast cancer invasion and metastasis. DEK has also been shown to maintain stemness [138] [139] and promotes mammosphere formation [138].

SSRP1 is known to be a subunit of the FACT complex which regulates transcription [131] . It also aids transcription by dissociating histone proteins from active sites [131] and acts as a transcriptional co-activator [132]. In terms of its roles in breast cancer, SSRP1 has been shown to be involved in cell proliferation, transformation , and maintaining stemness [142] [143].

PSIP1 promotes transcription by acting as a transcriptional co-activator [135]. Moreover, it regulates the expression of cell cycle genes and is involved in tumour migration and invasion in breast cancer [144].

Their roles in this project are investigated by the RNA interference technique, which uses shRNA (exogenous RNA) to target endogenous mRNA transcripts in order to achieve degradation and consequently a reduction in protein expression. This technique has the advantage of shRNA being constitutively expressed in the cells. In the single shRNA knock-down experiments, Figures 5.2- 5.4 show successful reduction of *DEK*, *SSRP1* and *PSIP1* expressions. However, these are not reflected in the protein levels of SSRP1 and PSIP1 p75 subunits. To understand the roles of these factors in this project better, I start by investigating their impacts

on the expression of *BCL11A*. Among them, knocking-down PSIP1 had no effect on *BCL11A*'s RNA expression but results in a partial reduction in its protein expression. These results suggest that PSIP1 was perhaps participating in regulating the translation of BCL11A. I then move on to investigate these factors' functional roles using cell proliferation and colony formation assays. Knocking-down these factors individually did not have any effect on assays but this is not enough to completely rule out the possibility of them regulating the expression of *BCL11A* or their having functional effects by working with each other.

To address this hypothesis, I continue the investigation by co-transfecting shRNA against two transcripts at the same time. Same as the results using single shRNA, the three factors are successfully knocked-down in both RNA and protein levels. Among these samples, the expression of *BCL11A* was also found to be reduced at both RNA and protein levels. However, only cells co-transfected with shRNAs against *SSRP1* and *PSIP1* did not affect *PPP2R2D*'s expression. *PPP2R2D* is a gene used as a control during RNA quantification and an example of housekeeping genes. These genes have functions that are important for cell survival and they are stably expressed in all biological systems [155]. For example, *PPP2R2D* is important for regulating a cell cycle check-point in mitosis [154]. Therefore, a reduction in this gene suggests that these proteins take part in regulating gene expressions on a global scale and this is also observed in cells transfected with DEK shRNA in Figure 5.7. On the other hand, no effect on the expression of *PPP2R2D* by knocking-down *SSRP1* and *PSIP1* suggests that these two proteins have potential roles in regulating the expression of *BCL11A*.

Interpreting results from Chapter 5 in the context of existing literature leads to my proposal of how SSRP1 and PSIP1 may regulate *BCL11A*'s expression. Firstly, the fact that knocking-down *PSIP1* alone has impact on *BCL11A*'s protein but not RNA expression suggests that PSIP1 has a role in regulating *BCL11A*'s translation. Secondly, PSIP1 is also thought to downregulate global gene expression. This is deduced from differences between the single knock-down of *PSIP1* and double knock-down of *PSIP1* and *SSRP1*. To quickly recap from Section 5.2.5.1, in both cases, *SSRP1* is absent and an effect on global gene expression shown by the

reduction in *PPP2R2D* expression is observed when PSIP1 is present. Thirdly, SSRP1 and PSIP1 are thought to regulate *BCL11A*'s expression together. This interpretation is based on results showing no effects on the expression of *BCL11A* at the RNA level when knocking-down *PSIP1* or *SSRP1* alone; contrast this to the double knock-down of both at the same time which results in a significant reduction in *BCL11A*'s expression. These suggest that SSRP1 and PSIP1 may take part in regulating *BCL11A*'s expression and that they may act as transcriptional co-activators. Furthermore, if PSIP1 does downregulate global gene expressions as suggested earlier, then there is the possibility that SSRP1 has also been downregulated. I therefore propose that SSRP1 may regulate *BCL11A*'s expression as a transcriptional co-activator while recruiting different transcription factors to those recruited by PSIP1.

6.2 Future directions

In this study, I have demonstrated that DEK, SSRP1 and PSIP1 are potential regulators of *BCL11A* using CRISPR-Cas9 and RIME techniques. However, there are several challenges that have to be addressed in the future. Improvements can also be made to the experiments to further extend our understanding on how these factors regulate the expression of *BCL11A* and their other roles in TNBC.

Firstly, as discussed, I have adapted the CRISPR-Cas9 system to study transcription at transcription active sites of *BCL11A*. The use of dCas9 has allowed me to avoid DNA double-strand breaks and use this protein as a tag for the RIME experiment that follows. Since the start of this project, a better version has been identified – APEX-Cas9. Peroxidase APEX2 is fused with dCas9 in this system and its major benefit is the quick proximal labelling of protein interactions [158], compared to the 24 hours or more wait using the Tet-On system in our current experimental design. Another advantage of this proximal protein labelling is that APEX2 will biotinylate proteins in close proximity, thus pulling down only those biotinylated by APEX2 and minimise potential noise signals during protein identification in comparison to crosslinking by formaldehyde. This can lead to further elimination

of non-specific proteins that may sit in close proximity but does not participate in regulating *BCL11A*'s expression. A third benefit of this APEX-Cas9 system will be the delivery of one plasmid that co-expresses both APEX-Cas9 and the gRNA. In comparison to our current system, combining the components of the CRISPR-Cas9 and this APEX system will allow us to speed up the process when establishing stable cell lines and mitigate challenges encountered in this project.

Secondly, I have demonstrated that RIME can be applied to identify previously unknown proteins: DEK, SSRP1, and PSIP1. To establish if these proteins specifically regulate *BCL11A*'s expression in the breast cancer context, this technique with the CRISPR-Cas9 system should be applied to other cell lines including normal cells, various TNBC and Luminal cell lines. Analysing these RIME datasets might even broaden our view on *BCL11A*'s regulation in TNBC by identifying more putative factors. For instance, the identification of *BCL11A* as oncogene in Lung Squamous Carcinoma [73] makes it interesting to investigate whether DEK, SSRP1, and PSIP1 also participate in regulating *BCL11A*'s expression in that context. This will further contribute to our understanding of the regulation of *BCL11A*, regardless of whether it is specific or context dependent. Although the identification of these three factors shed light on the regulation of *BCL11A*, the pathway and mechanism are currently unknown. These should be the focus in the future using techniques such as RNA-sequencing to detect changes in gene expressions and mass spectrometry to compare protein expressions in the various conditions mentioned.

Thirdly, I validated the expressions of DEK, SSRP1, and PSIP1 in various Luminal and TNBC cell lines. Aside from these cell lines, expressions of DEK, SSRP1, and PSIP1 should also be investigated in other breast cancer cell lines and PDXs. Moreover, their impacts on *BCL11A*'s expression have been determined and functional studies attempted. Again, these impacts should be investigated in other breast cancer cell lines, PDXs, and organoids. Additionally, more functional studies have to be done to conclude if DEK, SSRP1, and PSIP1 have any effects functionally. It would also be interesting to see if these factors bind at other genes by performing ChIP-sequencing. This procedure can potentially help us to

understand whether these factors bind at the same sites or act through other mechanisms such as recruiting other DNA regions, e.g. enhancers, to the promoter region of *BCL11A* and form DNA looping; or to identify potential downstream genes regulated by these factors or by our oncogene.

The impact of *BCL11A*'s expression was investigated using shRNA knock-down. One criticism on the experimental design could be about the selection marker used in the combination shRNA knock-down experiments. To have confidence that the established cell line consists of plasmids that target different factors, several selection markers such as resistance genes for different antibiotics and different fluorescence markers should be used. Furthermore, the results from shRNA knock-down in this project should be validated in other breast cancer cell lines, PDXs, and organoids; and investigated further in mouse models. Another challenge I have encountered after establishing stable cell lines to co-express two shRNAs was the reduction in growth rate. This could be tackled in the future using inducible shRNA plasmids and extending the project to further understand the effects of these three proteins by knocking them all down at the same time.

Therapeutic potential implied by my current results can be further investigated using existing drugs, such as CBL0137 which affects *SSRP1* functions [159] [160] [161] [162] [163] [164] and is currently administered as a therapy for hematological malignancies [165] and undergoing clinical trial for treating small cell lung cancer [163]. CBL0137 is a Curaxcin-based drug that acts by changing the topology of DNA through intercalation [160]. The consequence of this is that the FACT complex, which consists of SSRP1 and SPT16H subunits, becomes unable to bind to DNA and transcribe genes such as the Nuclear Factor kappa-light-chain-enhancer of activated B cells (NF-κB) [160] [166]. The inhibition of the expression of NF-κB induces cell apoptosis [166]. At the same time, CBL0137 also promotes the formation of a complex with the FACT complex and Caesin kinase 2 (CK2), and phosphorylates the Serine at position 392 of p53 which then activates p53 [148]. This simultaneous mechanism of CBL0137 results in a reduction in tumour growth and induces cell apoptosis.

Lastly, I have attempted to establish if these three proteins form a protein complex. However, this has proven to be technical challenging and I was unable to finish optimisations due to time constraints. The Co-IP experiment could be further optimised and the investigation could be extended to multiple breast cancer cell lines and tumour samples.

6.3 Conclusion

To conclude, I established a novel system to study transcriptional regulation of *BCL11A* in TNBC using the CRISPR-Cas9 system and identify three factors as potential regulators using RIME proteomics. Adaptation of the CRISPR-Cas9 system in this project involves using dCas9 to not generate DNA double-strand breaks but as a tag for pulling down itself with proteins in close proximity at our sites of interests. Analysis of these proteins with the RIME proteomics dataset using multiple parameters and criteria leads to the identification of DEK, SSRP1, and PSIP1. From there, their effects on the expression of *BCL11A* were investigated using shRNA knock-down. Combining results from different knock-down experiments, I propose that SSRP1 and PSIP1 play a bigger role in regulating *BCL11A*'s expression through different mechanisms. These results imply that DEK, SSRP1, and PSIP1 could be potential drug targets when developing therapeutic targets against *BCL11A* specifically for TNBC. This implication is particularly important when seen as a potential answer to the clinical challenges of targeting *BCL11A* due to its general importance in celluar functions and development outside of the TNBC context, as previously mentioned in the original aims and rationales behind this project.

REFERENCES

[1] Cancer Research UK (2015), "Breast cancer incidence (invasive) statistics," https://www.cancerresearchuk.org/health-professional/cancer-statistics /statistics-by-cancer-type/breast-cancer/incidence-invasive#heading-Zero. [Accessed on 26.03.2016]

[2] Cancer Research UK (2017), "What is breast cancer?," https://www.cancerresearchuk.org/about-cancer/breast-cancer/about. [Accessed on 14.01.2018]

[3] P. Boyle (2012) "Triple-negative breast cancer: epidemiological considerations and recommendations," *Annals of Oncology,* vol. 23, no. suppl_6, pp. vi7-vi12

[4] J. A. Bray F., Ferlay J., Soerjomataram et al, (2018) "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians,* vol. 686, pp. 394-424

[5] M. G. Marmot, D. G. Altman, D. A. Cameron et al (2013), "The benefits and harms of breast cancer screening: an independent review," *British journal of cancer,* vol. 108, no. 11, pp. 2205-2240.

[6] Breast Cancer care (2019), "Facts and statistics 2019," https://www.breastcancercare.org.uk/about-us/media/facts-statistics. [Accessed on 15.03.2019]

[7] K. M. O'Brien, S. R. Cole, C. K. Tse et al (2010), "Intrinsic breast tumor subtypes, race, and long-term survival in the carolina breast cancer study," *Clinical Cancer Research,* vol. 16, no, 24, pp. 6100-6110

[8] C. Brisken, S. Park, T. Vass et al (1998), "A paracrine role for the epithelial progesterone receptor in mammary gland development," *Proceedings of the National Academy of Sciences,* vol. 95, no. 9, p. 5076,

[9] K. P. Ormandy C.J., Camus A., Barra J. et al (1997), "Null mutation of the prolactin receptor gene produces multiple reproductive defects in the mouse.," *Genes & Development,* vol. 11, pp. 167-178

[10] W. G. Juergens, F. E. Stockdale, Y. J. Topper et al (1965), "Hormone-dependent differentiation of mammary gland in vitro," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 54, no. 2, pp. 629-634

[11] C. Brisken (2002), "Hormonal Control of Alveolar Development and Its Implications for Breast Carcinogenesis," *Journal of Mammary Gland Biology and Neoplasia,* vol. 7, no. 1, pp. 39-48

[12] O. M. Inada K.,Tominaga T.,Toi M et al (1996), "Progesterone-stimulated growth of mammary carcinomas induced by 7,12-dimethylbenz[a]anthracene in neonatally androgenized rats.," *European Journal of Surgical Oncology (EJSO),* vol. 22, no. 6, pp. 583-587

[13] J. M. Williams and C. W. Daniel (1983), "Mammary ductal elongation: Differentiation of myoepithelium and basal lamina during branching morphogenesis," *Developmental Biology,* vol. 97, no. 2, pp. 274-290,

[14] H. H. Traurig (1967), "Cell proliferation in the mammary gland during late pregnancy and lactation," *The anatomical record.,* vol. 157, no. 3, pp. 489-503

[15] M. Shackleton, F. Vaillant, K. J. Simpson et aal (2006), "Generation of a functional mammary gland from a single stem cell," *Nature,* vol. 439, no. 7072, pp. 84-88

[16] C. J. Watson and W. T. Khaled (2008), "Mammary development in the embryo and adult: a journey of morphogenesis and commitment," *Development,* vol. 135, no. 6, p. 995

[17] J. O'Brien, H. Martinson, C. Durand-Rougely et al (2011), "Macrophages are crucial for epithelial cell death and adipocyte repopulation during mammary gland involution," *Development,* vol. 139, no. 2, pp. 269-275

[18] T. J. Sargeant, B. Lloyd-Lewis, H. K. Resemann et al (2014), "Stat3 controls cell death during mammary gland involution by regulating uptake of milk fat globules and lysosomal membrane permeabilization," *Nature cell biology,* vol. 16, no. 11, pp. 1057-1068

[19] P. A. Kreuzaler, A. D. Staniszewska, W. Li et al (2011), “Stat3 controls lysosomal-mediated cell death in vivo,” *Nature Cell Biology,* vol. 13, p. 303

[20] T. A. Ince, A. L. Richardson, G. W. Bell et al (2007), “Transformation of Different Human Breast Epithelial Cell Types Leads to Distinct Tumor Phenotypes,” *Cancer Cell,* vol. 12, no. 2, pp. 160-170

[21] A. H. Sims, A. Howell, S. J. Howell et al (2007), “Origins of breast cancer subtypes and therapeutic implications,” *Nature Clinical Practice Oncology,* vol. 4, p. 516

[22] Watson C. J. and Khaled W. (2008), "Mammary gland development in the embryo and adult: a journey of morphogenesis and commitment," *Development*, vol. 135, no. 995-1003

[23] S. Chumsri, T. Howes, T. Bao et al (2011), “Aromatase, aromatase inhibitors, and breast cancer,” *The Journal of Steroid Biochemistry and Molecular Biology,* vol. 125, no. 1, pp. 13-22

[24] E. H. B. C. C. Group (2003), “Body Mass Index, Serum Sex Hormones, and Breast Cancer Risk in Postmenopausal Women,” *JNCI: Journal of the National Cancer Institute,* vol. 95, no. 16, pp. 1218-1226

[25] T. Sørlie, R. Tibshirani, J. Parker et al (2003), “Repeated observation of breast tumor subtypes in independent gene expression data sets,” *Proceedings of the National Academy of Sciences,* vol. 100, no. 14, p. 8418

[26] S. Tyanova, R. Albrechtsen, P. Kronqvist et al (2016), “Proteomic maps of breast cancer subtypes,” *Nature Communications,* vol. 7, p. 10259

[27] C. M. Perou, T. Sørlie, M. B. Eisen et al (2000), “Molecular portraits of human breast tumours,” *Nature,* vol. 406, no. 6797, pp. 747-752

[28] R. Sabatier, P. Finetti, A. Guille et al (2014), “Claudin-low breast cancers: clinical, pathological, molecular and prognostic characterization,” *Molecular cancer,* vol. 13, p. 228

[29] T. Sørlie, C. M. Perou, R. Tibshirani et al (2001), “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical

implications," *Proceedings of the National Academy of Sciences,* vol. 98, no. 19, p. 10869

[30] C. Curtis, S. P. Shah, S.-F. Chin et al (2012), "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature,* vol. 486, p. 346

[31] K. Tomczak, P. Czerwińska and M. Wiznerowicz (2015), "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemporary oncology (Poznan, Poland),* vol. 19, no. 1A, pp. A68-A77

[32] E. A. Rakha, S. E. Elsheikh, M. A. Aleskandarany et al (2009), "Triple-Negative Breast Cancer: Distinguishing between Basal and Nonbasal Subtypes," *Clinical Cancer Research,* vol. 15, no. 7, p. 2302

[33] R. Dent, M. Trudeau, K. I. Pritchard et al (2007), "Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence," *Clinical Cancer Research,* vol. 13, no. 15, p. 4429

[34] J. B. Arnes, J.-S. Brunet, I. Stefansson et al (2005), "Placental Cadherin and the Basal Epithelial Phenotype of BRCA1-Related Breast Cancer," *Clinical Cancer Research,* vol. 11, no. 11, p. 4003

[35] M. C. U. Cheang, D. Voduc, C. Bajdik et al (2008), "Basal-Like Breast Cancer Defined by Five Biomarkers Has Superior Prognostic Value than Triple-Negative Phenotype," *Clinical Cancer Research,* vol. 14, no. 5, p. 1368

[36] P. T. Simpson, L. M. Da Silva and S. R. Lakhani (2007), *In Situ Carcinoma-Can We Predict which Patient Will Come Back with a Recurrence?,* vol. 12, pp. 409-411

[37] P. Boström, M. Söderström, T. Palokangas et al (2009), "Analysis of cyclins A, B1, D1 and E in breast cancer in relation to tumour grade and other prognostic factors," *BMC Research Notes,* vol. 2, no. 1, p. 140

[38] D. Voduc, T. O. Nielsen, M. C. Cheang et al (2008), "The combination of high cyclin E and Skp2 expression in breast cancer is associated with a poor prognosis and the basal phenotype," *Human Pathology,* vol. 39, no. 10, pp. 1431-1437

[39] T. C. G. A. Network, D. C. Koboldt, R. S. Fulton et al (2012), "Comprehensive molecular portraits of human breast tumours," *Nature,* vol. 490, p. 61

[40] A. P. Subhawong, T. Subhawong, H. Nassar et al (2009), "Most basal-like breast carcinomas demonstrate the same Rb-/p16+ immunophenotype as the HPV-related poorly differentiated squamous cell carcinomas which they resemble morphologically," *The American journal of surgical pathology,* vol. 33, no. 2, pp. 163-175

[41] R. D. Chacón and M. V. Costanzo (2010), "Triple-negative breast cancer," *Breast Cancer Research,* vol. 12, no. 2, p. S3

[42] D. Sarrió, S. M. Rodriguez-Pinilla, D. Hardisson et al (2008), "Epithelial-Mesenchymal Transition in Breast Cancer Relates to the Basal-like Phenotype," *Cancer Research,* vol. 68, no. 4, p. 989

[43] C. K. Anders and L. A. Carey (2009), "Biology, Metastatic Patterns, and Treatment of Patients with Triple-Negative Breast Cancer," *Clinical Breast Cancer,* vol. 9, pp. S73-S81

[44] V. G. Abramson, B. D. Lehmann, T. J. Ballinger et al (2015), *Subtyping of triple-negative breast cancer: Implications for therapy,* vol. 121, John Wiley and Sons Inc., 2015, pp. 8-16.

[45] D. Cunningham, Y. Humblet, S. Siena et al (2004), "Cetuximab Monotherapy and Cetuximab plus Irinotecan in Irinotecan-Refractory Metastatic Colorectal Cancer," *New England Journal of Medicine,* vol. 351, no. 4, pp. 337-345

[46] E. Van Cutsem, C.-H. Köhne, E. Hitre et al (2009), "Cetuximab and Chemotherapy as Initial Treatment for Metastatic Colorectal Cancer," *New England Journal of Medicine,* vol. 360, no. 14, pp. 1408-1417

[47] J. Kurai, H. Chikumi, K. Hashimoto et al (2007), "Antibody-Dependent Cellular Cytotoxicity Mediated by Cetuximab against Lung Cancer Cell Lines," *Clinical Cancer Research,* vol. 13, no. 5, p. 1552, 1 3 2007.

[48] R. Pirker, J. R. Pereira, A. Szczesna et al (2009), "Cetuximab plus chemotherapy in patients with advanced non-small-cell lung cancer

(FLEX): an open-label randomised phase III trial," *The Lancet,* vol. 373, no. 9674, pp. 1525-1531

[49] H. Q. Xiong, A. Rosenberg, A. LoBuglio et al (2004), "Cetuximab, a Monoclonal Antibody Targeting the Epidermal Growth Factor Receptor, in Combination With Gemcitabine for Advanced Pancreatic Cancer: A Multicenter Phase II Trial," *Journal of Clinical Oncology,* vol. 22, no. 13, pp. 2610-2616

[50] F. Robert, M. P. Ezekiel, S. A. Spencer et al (2001), "Phase I Study of Anti–Epidermal Growth Factor Receptor Antibody Cetuximab in Combination With Radiation Therapy in Patients With Advanced Head and Neck Cancer," *Journal of Clinical Oncology,* vol. 19, no. 13, pp. 3234-3243

[51] T. O. Nielsen, F. D. Hsu, K. Jensen et al (2004), "Immunohistochemical and Clinical Characterization of the Basal-Like Subtype of Invasive Breast Carcinoma," *Clinical Cancer Research,* vol. 10, no. 16, p. 5367

[52] L. A. Carey, H. S. Rugo, P. K. Marcom et al (2012), "TBCRC 001: randomized phase II study of cetuximab in combination with carboplatin in stage IV triple-negative breast cancer," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* vol. 30, no. 21, pp. 2615-2623

[53] S. Modi, G. D'Andrea, L. Norton et al (2006), "A Phase I Study of Cetuximab/Paclitaxel in Patients with Advanced-Stage Breast Cancer," *Clinical Breast Cancer,* vol. 7, no. 3, pp. 270-277

[54] A. M. Hosey, J. J. Gorski, M. M. Murray et al (2007), "Molecular Basis for Estrogen Receptor α Deficiency in BRCA1-Linked Breast Cancer," *JNCI: Journal of the National Cancer Institute,* vol. 99, no. 22, pp. 1683-1694

[55] J. J. Gorski, C. R. James, J. E. Quinn et al (2010), "BRCA1 transcriptionally regulates genes associated with the basal-like phenotype in breast cancer," *Breast Cancer Research and Treatment,* vol. 122, no. 3, pp. 721-731

[56] Hudisa C.A. and Gianni L. (2011), "Triple-Negative Breast Cancer: An Unmet Medical Need," *The Oncologist,* vol. 16, pp. 1-11

[57] P. Liu, J. R. Keller, M. Ortiz et al (2003), "Bcl11a is essential for normal lymphoid development," *Nature Immunology,* vol. 4, p. 525

[58] H. Liu, G. C. Ippolito, J. K. Wall et al (2006), "Functional studies of BCL11A: characterization of the conserved BCL11A-XL splice variant and its interaction with BCL6 in nuclear paraspeckles of germinal center B cells," *Molecular Cancer,* vol. 5, no. 1, p. 18

[59] E. Satterwhite, T. Sonoki, T. G. Willis et al (2001), "TheBCL11 gene family: involvement of BCL11A in lymphoid malignancies," *Blood,* vol. 98, no. 12, p. 3413

[60] Z. Chen, H.-y. Luo, M. H. Steinberg et al (2009), "BCL11A represses HBG transcription in K562 cells," *Blood Cells, Molecules, and Diseases,* vol. 42, no. 2, pp. 144-149

[61] M. Uda, R. Galanello, S. Sanna et al (2008), "Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of β-thalassemia," *Proceedings of the National Academy of Sciences,* vol. 105, no. 5, p. 1620

[62] K. Pulford, A. H. Banham, L. Lyne et al (2006), "The BCL11AXL transcription factor: its distribution in normal and malignant tissues and use as a marker for plasmacytoid dendritic cells," *Leukemia,* vol. 20, p. 1439

[63] V. G. Sankaran, T. F. Menne, J. Xu et al (2008), "Human Fetal Hemoglobin Expression Is Regulated by the Developmental Stage-Specific Repressor & BCL11A;," *Science,* vol. 322, no. 5909, p. 1839

[64] V. G. Sankaran, J. Xu and S. H. Orkin (2010), "Transcriptional silencing of fetal hemoglobin by BCL11A," *Annals of the New York Academy of Sciences,* vol. 1202, no. 1, pp. 64-68

[65] J. Xu, V. G. Sankaran, M. Ni et al (2010), "Transcriptional silencing of γ-globin by BCL11A involves long-range interactions and cooperation with SOX6," *Genes & development,* vol. 24, no. 8, pp. 783-798,

[66] M. A. Weniger, K. Pulford, S. Gesk et al (2006), "Gains of the proto-oncogene BCL11A and nuclear accumulation of BCL11AXL protein are

frequent in primary mediastinal B-cell lymphoma," *Leukemia,* vol. 20, p. 1880

[67] Y. Yu, J. Wang, W. Khaled et al (2012), "Bcl11a is essential for lymphoid development and negatively regulates p53," *The Journal of Experimental Medicine,* vol. 209, no. 13, p. 2467

[68] N. Liu, J. Zhang and C. Ji (2013), "The emerging roles of Notch signaling in leukemia and stem cells," *Biomarker Research,* vol. 1, no. 1, p. 23

[69] K. Jawaid, K. Wahlberg, S. L. Thein et al (2010), "Binding patterns of BCL11A in the globin and GATA1 loci and characterization of the BCL11A fetal hemoglobin locus," *Blood Cells, Molecules, and Diseases,* vol. 45, no. 2, pp. 140-146

[70] C. Dias, S. Estruch, S. Graham et al (2016), "BCL11A Haploinsufficiency Causes an Intellectual Disability Syndrome and Dysregulates Transcription," *The American Journal of Human Genetics,* vol. 99, no. 2, pp. 253-274

[71] A. Basak, M. Hancarova, J. C. Ulirsch et al (2015), "BCL11A deletions result in fetal hemoglobin persistence and neurodevelopmental alterations," *The Journal of Clinical Investigation,* vol. 125, no. 6, pp. 2363-2368

[72] K. A. Lazarus, F. Hadi, E. Zambon et al (2018), "BCL11A interacts with SOX2 to control the expression of epigenetic regulators in lung squamous carcinoma," *Nature Communications,* vol. 9, no. 1, p. 3327

[73] W. T. Khaled, S. Choon Lee, J. Stingl et al (2015), "BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells," *Nature Communications,* vol. 6, p. 5987

[74] M. Ehrlich (2009), "DNA hypomethylation in cancer cells," *Epigenomics,* vol. 1, no. 2, pp. 239-259

[75] I. Grummt and C. S. Pikaard (2003), "Epigenetic silencing of RNA polymerase I transcription," *Nature Reviews Molecular Cell Biology,* vol. 4, p. 641

[76] J. M. G. Vilar and S. Leibler (2003), "DNA Looping and Physical Constraints on Transcription Regulation," *Journal of Molecular Biology,vol.* 331, vol. 5, pp. 981-989

[77] Mariana Ruiz Villarreal, "Transcription," https://www.ck12.org/biology/transcription/lesson/Transcription-of-DNA-to-RNA-BIO/. [Accessed on 18.12.2018]

[78] A. Kulyyassov, G. Zhubanova, E. Ramanculov et al (2015), "Proximity Utilizing Biotinylation of Nuclear Proteins in vivo," *Central Asian Journal of Global Health,* vol. 3

[79] W. Wang, J. Sun, M. Nimtz et al (2003), "Protein identification from two-dimensional gel electrophoresis analysis of Klebsiella pneumoniae by combined use of mass spectrometry data and raw genome sequences," *Proteome science,* vol. 1, p. 6

[80] H. Mohammed and J. S. Carroll (2013), "Approaches for assessing and discovering protein interactions in cancer," *Molecular cancer research : MCR,* vol. 11, no. 11, pp. 1295-1302.

[81] E. Klenova, I. Chernukhin, T. Inoue (2002), "Immunoprecipitation techniques for the analysis of transcription factor complexes," *Methods,* vol. 26, no. 3, pp. 254-259

[82] H. Mohammed, C. Taylor, G. D. Brown et al (2016), "Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) for analysis of chromatin complexes," *Nature Protocols,* vol. 11, p. 316

[83] K. C. Duong-Ly and S. B. Gabelli (2015), "Chapter Two - Affinity Purification of a Recombinant Protein Expressed as a Fusion with the Maltose-Binding Protein (MBP) Tag," *Methods in Enzymology*, vol. 559, J. R. Lorsch,, Academic Press, 2pp. 17-26.

[84] P. A. Latos, A. Goncalves, D. Oxley et al (2015), "Fgf and Esrrb integrate epigenetic and transcriptional networks that regulate self-renewal of trophoblast stem cells," *Nature Communications,vol.* 6, p. 7776, 24 7 2015.

[85] S. Stelloo, E. Nevedomskaya, Y. Kim et al (2017), "Endogenous androgen receptor proteomic profiling reveals genomic subcomplex involved in prostate tumorigenesis," *Oncogene,* vol. 37, p. 313

[86] S. J. Barfeld, A. Urbanucci, H. M. Itkonen et al (2017), "c-Myc Antagonises the Transcriptional Activity of the Androgen Receptor in Prostate Cancer Affecting Key Gene Networks," *EBioMedicine,* vol. 18, pp. 83-93

[87] H. Mohammed, C. D'Santos, A. Serandour et al (2013), "Endogenous Purification Reveals GREB1 as a Key Estrogen Receptor Regulatory Factor," *Cell Reports,* vol. 3, no. 2, pp. 342-349

[88] E. Deltcheva, K. Chylinski, C. M. Sharma et al (2011), "CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III," *Nature,* vol. 471, p. 602

[89] J. D. Sander and J. K. Joung (2014), "CRISPR-Cas systems for editing, regulating and targeting genomes," *Nature Biotechnology,* vol. 32, p. 347

[90] C. Pourcel, G. Vergnaud and I. Grissa (2007), "CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats," *Nucleic Acids Research,* vol. 35, no. suppl_2, pp. W52-W57

[91] B. Wiedenheft, E. van Duijn, J. B. Bultema et al (2011), "RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions," *Proceedings of the National Academy of Sciences,* vol. 108, no. 25, p. 10092

[92] E. Semenova, M. M. Jore, K. A. Datsenko et al (2011), "Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence," *Proceedings of the National Academy of Sciences,* vol. 108, no. 25, p. 10098

[93] P. Mali, L. Yang, K. M. Esvelt et al (2013), "RNA-Guided Human Genome Engineering via Cas9," *Science,* vol. 339, no. 6121, p. 823

[94] L. Cong, F. A. Ran, D. Cox et al (2013), "Multiplex Genome Engineering Using CRISPR/Cas Systems," *Science,* vol. 339, no. 6121, p. 819

[95] E. C. Friedberg (2003), "DNA damage and repair," *Nature,* vol. 421, no. 6921, pp. 436-440

[96] E. Weterings and D. J. Chen (2008), "The endless tale of non-homologous end-joining," *Cell Research,* vol. 18, p. 114

[97] L. Davis, Y. Zhang and N. Maizels (2018), "Chapter Five - Assaying Repair at DNA Nicks," *Methods in Enzymology*, vol. 601, pp. 71-89.

[98] C. Mussolino, R. Morbitzer, F. Lütge et al (2011), "A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity," *Nucleic Acids Research,* vol. 39, no. 21, pp. 9283-9293

[99] D. Carroll (2011), "Genome Engineering With Zinc-Finger Nucleases," *Genetics,* vol. 188, no. 4, p. 773

[100] M. Jinek, K. Chylinski, I. Fonfara et al (2012), "A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity," *Science,* vol. 337, no. 6096, p. 816

[101] Zhang lab, "Guide Design Resources," https://zlab.bio/guide-design-resources. [Accessed 19.01.2015]

[102] S.-E. Shin, J.-M. Lim, H. G. Koh et al (2016), "CRISPR/Cas9-induced knockout and knock-in mutations in Chlamydomonas reinhardtii," *Scientific Reports,* vol. 6, p. 27810

[103] R. Platt, S. Chen, Y. Zhou et al (2014), "CRISPR-Cas9 Knockin Mice for Genome Editing and Cancer Modeling," *Cell,* vol. 159, no. 2, pp. 440-455

[104] S. Konermann, M. D. Brigham, A. E. Trevino, et al (2014), "Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex," *Nature,* vol. 517, p. 583

[105] J. Black, A. Adler, H.-G. Wang et al (2016), "Targeted Epigenetic Remodeling of Endogenous Loci by CRISPR/Cas9-Based Transcriptional Activators Directly Converts Fibroblasts to Neuronal Cells," *Cell Stem Cell,* vol. 19, no. 3, pp. 406-414

[106] D. G. Ousterout, A. M. Kabadi, P. I. Thakore et al (2015), "Multiplex CRISPR/Cas9-based genome editing for correction of dystrophin mutations

that cause Duchenne muscular dystrophy," *Nature Communications,* vol. 6, p. 6244

[107] T. O. Auer, K. Duroure, A. De Cian et al (2014), "Highly efficient CRISPR/Cas9-mediated knock-in in zebrafish by homology-independent DNA repair," *Genome research,* vol. 24, no. 1, pp. 142-153

[108] F. Ran, P. Hsu, C.-Y. Lin et al (2013), "Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity," *Cell,* vol. 154, no. 6, pp. 1380-1389

[109] L. Gilbert, M. Larson, L. Morsut et al (2013), "CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes," *Cell,* vol. 154, no. 2, pp. 442-451

[110] A. Fire, S. Xu, M. K. Montgomery et al (1998), "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans," *Nature,* vol. 391, no. 6669, pp. 806-811

[111] S. M. Hammond, E. Bernstein, D. Beach et al (2000), "An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells," *Nature,* vol. 404, no. 6775, pp. 293-296

[112] T. R. Brummelkamp, R. Bernards and R. Agami (2002), "A System for Stable Expression of Short Interfering RNAs in Mammalian Cells," *Science,* vol. 296, no. 5567, p. 550

[113] H. Xia, Q. Mao, H. L. Paulson et al (2002), "siRNA-mediated gene silencing in vitro and in vivo," *Nature Biotechnology,* vol. 20, p. 1006

[114] T. R. Brummelkamp, R. Bernards and R. Agami (2002), "Stable suppression of tumorigenicity by virus-mediated RNA interference," *Cancer Cell,* vol. 2, no. 3, pp. 243-247

[115] G. J. Hannon (2002), "RNA interference," *Nature,* vol. 418, no. 6894, pp. 244-251

[116] UCSC genome database browser, "Human BCL11A gene locus," http://genomeeuro.ucsc.edu/cgibin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode

=0&nonVirtPosition=&position=chr2%3A60457192605551 54&hgsid=231827849_Re7AHQjprqgm7EzwUbvaJDZfvGKY.
[Accessed on 19.01.2015]

[117] Sigma, "MISSION shRNA", https://www.sigmaaldrich.com/life-science/functional-genomics-and-rnai/shrna/individual-genes.html [Accessed on 10.04.2018]

[118] T.-Y. Kuo, C.-J. Hong and Y.-P. Hsueh (2009), "Bcl11A/CTIP1 regulates expression of DCC and MAP1b in control of axon branching and dendrite outgrowth," *Molecular and Cellular Neuroscience,* vol. 42, no. 3, pp. 195-207

[119] W. J. Kent, C. W. Sugnet, T. S. Furey, et al (2002), "The human genome browser at UCSC," *Genome research,* vol. 12, no. 6, pp. 996-1006

[120] M. P. Creyghton, A. W. Cheng, G. G. Welstead et al (2010), "Histone H3K27ac separates active from poised enhancers and predicts developmental state," *Proceedings of the National Academy of Sciences,* vol. 107, no. 50, pp. 21931-21936.

[121] P. N. Cockerill (2011), *Structure and function of active chromatin and DNase i hypersensitive sites,* vol. 278, pp. 2182-2210.

[122] K. J. Chavez, S. V. Garimella and S. Lipkowitz (2010), "Triple negative breast cancer cell lines: one tool in the search for better treatment of triple negative breast cancer," *Breast disease,* vol. 32, no. 1-2, pp. 35-48

[123] Y. Wang, J. Wang, A. Devaraj et al (2014), "Suicidal Autointegration of Sleeping Beauty and piggyBac Transposons in Eukaryotic Cells," *PLOS Genetics,* vol. 10, no. 3, pp. e1004103-

[124] Cambridge Bioscience, "PiggyBac Transposon Vector Construction," https://www.bioscience.co.uk/products/piggybac-transposon-vector-construction. [Accessed on 24.03.2018]

[125] J. Szulc, M. Wiznerowicz, M.-O. Sauvain et al (2006), "A versatile tool for conditional gene expression and knockdown", *Nature Mathods*, vol. 3, pp. 109-116.

[126] D. Mellacheruvu, Z. Wright, A. L. Couzens et al (2013), "The CRAPome: a contaminant repository for affinity purification–mass spectrometry data," *Nature Methods,* vol. 10, p. 730

[127] T. Waldmann, C. Eckerich, M. Baack et al (2002), The Ubiquitous Chromatin Protein DEK Alters the Structure of DNA by Introducing Positive Supercoils, vol. 277, pp. 24988-24994.

[128] V. Alexiadis, T. Waldmann, J. Andersen et al (2000), "The protein encoded by the proto-oncogene DEK changes the topology of chromatin and reduces the efficiency of DNA replication in a chromatin-specific manner," *Genes & development,* vol. 14, no. 11, pp. 1308-1312

[129] L. M. M. Soares, K. Zanier, C. Mackereth et al (2006), "Intron Removal Requires Proofreading of U2AF/3' Splice Site Recognition by DEK," *Science,* vol. 312, no. 5782, p. 1961

[130] G. Orphanides, W.-H. Wu, W. S. Lane et al (1999), "The chromatin-specific transcription elongation factor FACT comprises human SPT16 and SSRP1 proteins," *Nature,* vol. 400, no. 6741, pp. 284-288

[131] S. X. Zeng, M.-S. Dai, D. M. Keller et al (2002), "SSRP1 functions as a co-activator of the transcriptional activator p63," *The EMBO journal,* vol. 21, no. 20, pp. 5487-5497

[132] E. Meshorer and T. Misteli (2006), "Chromatin in pluripotent embryonic stem cells and differentiation," *Nature Reviews Molecular Cell Biology,* vol. 7, p. 540

[133] D. P. Singh, N. Ohguro, T. Kikuchi et al (2000), "Lens Epithelium-Derived Growth Factor: Effects on Growth and Survival of Lens Epithelial Cells, Keratinocytes, and Fibroblasts," *Biochemical and Biophysical Research Communications,* vol. 267, no. 1, pp. 373-381

[134] L. T. Chylack, L. Fu, R. Mancini et al (2004), "Lens epithelium-derived growth factor (LEDGF/p75) expression in fetal and adult human brain," *Experimental Eye Research,* vol. 79, no. 6, pp. 941-948

[135] H. M. Marshall, K. Ronen, C. Berry et al (2007), "Role of PSIP1/LEDGF/p75 in Lentiviral Infectivity and Integration Targeting," *PLOS ONE,* vol. 2, no. 12, pp. e1340-

[136] G. Ying and Y. Wu (2015), DEK: A novel early screening and prognostic marker for breast cancer, *Molecular Medicine Reports*, vol. 12, p. 7491-7495

[137] L. M. Privette Vinnedge, R. McClaine, P. K. Wagh et al (2011), "The human DEK oncogene stimulates β-catenin signaling, invasion and mammosphere formation in breast cancer," *Oncogene,* vol. 30, p. 2741

[138] Y. Yang, M. Gao, Z. Lin et al (2017), "DEK promoted EMT and angiogenesis through regulating PI3K/AKT/mTOR pathway in triple-negative breast cancer," *Oncotarget,* vol. 8, no. 58, pp. 98708-98722

[139] K. Attwood, D. Fleyshman, L. Prendergast et al (2017), "Prognostic value of histone chaperone FACT subunits expression in breast cancer," *Breast cancer (Dove Medical Press),* vol. 9, pp. 301-311

[140] D. Fleyshman, L. Prendergast, A. Safina et al (2017), "Level of FACT defines the transcriptional landscape and aggressive phenotype of breast cancer cells," *Oncotarget,* vol. 8, no. 13, pp. 20525-20542

[141] C. Morrison, A. R. Omilian, K. V. Gurova et al (2013), "Level of SSRP1 in Cancer as a Prognostic Marker of Aggressive Disease," *American Journal of Clinical Pathology,* vol. 140, no. suppl_1, pp. A152-A152

[142] H. Garcia, J. Miecznikowski, A. Safina et al (2013), "Facilitates Chromatin Transcription Complex Is an "Accelerator" of Tumor Transformation and Potential Marker and Target of Aggressive Cancers," *Cell Reports,* vol. 4, no. 1, pp. 159-173

[143] M. von Lindern, M. Fornerod, S. van Baal et al (1992), "The translocation (6;9), associated with a specific subtype of acute myeloid leukemia, results in the fusion of two genes, dek and can, and the expression of a chimeric, leukemia-specific dek-can mRNA.," *Molecular and Cellular Biology,* vol. 12, no. 4, p. 1687

[144] L. M. Privette Vinnedge, N. M. Benight, P. K. Wagh et al (2014), "The DEK oncogene promotes cellular proliferation through paracrine Wnt signaling in Ron receptor-positive breast cancers," *Oncogene,* vol. 34, p. 2325

[145] B. C.-M. Tan, C.-T. Chien, S. Hirose et al (2006), "Functional cooperation between FACT and MCM helicase facilitates initiation of chromatin DNA replication," *The EMBO Journal,* vol. 25, no. 17, pp. 3975-3985

[146] D. M. Keller, X. Zeng, Y. Wang et al (2001), "A DNA Damage–Induced p53 Serine 392 Kinase Complex Contains CK2, hSpt16, and SSRP1," *Molecular Cell,* vol. 7, no. 2, pp. 283-292

[147] R. van Nuland, F. M. van Schaik, M. Simonis et al (2013), "Nucleosomal DNA binding drives the recognition of H3K36-methylated nucleosomes by the PSIP1-PWWP domain," *Epigenetics & chromatin,* vol. 6, no. 1, p. 12

[148] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt et al (2003), "STRING: a database of predicted functional associations between proteins," *Nucleic acids research,* vol. 31, no. 1, pp. 258-261

[149] A. P. Lopez, J. R. Kugelman, J. Garcia-Rivera et al (2016), "The Structure-Specific Recognition Protein 1 Associates with Lens Epithelium-Derived Growth Factor Proteins and Modulates HIV-1 Replication," *Journal of molecular biology,* vol. 428, no. 14, pp. 2814-2831

[150] S. Delaunay, F. Rapino, L. Tharun et al (2016), "Elp3 links tRNA modification to IRES-dependent translation of LEF1 to sustain metastasis in breast cancer," *The Journal of Experimental Medicine,* vol. 213, no. 11, p. 2503

[151] A. T. Yarnell, S. Oh, D. Reinberg and S. J. Lippard (2001), "Interaction of FACT, SSRP1, and the High Mobility Group (HMG) Domain of SSRP1 with DNA Damaged by the Anticancer Drug Cisplatin," *Journal of Biological Chemistry,* vol. 276, no. 28, pp. 25736-25741

[152] S. Mochida, S. L. Maslen, M. Skehel et al (2010), "Greatwall Phosphorylates an Inhibitor of Protein Phosphatase 2A That Is Essential for Mitosis," *Science*, vol. 330, no. 6011, p. 1670

[153] E. Eisenberg and E. Y. Levanon (2013), "Human housekeeping genes, revisited," *Trends in Genetics,* vol. 29, no. 10, pp. 569-574

[154] K. V. Sitwala, K. Adams and D. M. Markovitz (2002), "YY1 and NF-Y binding sites regulate the transcriptional activity of the dek and dek-can promoter," *Oncogene,* vol. 21, p. 8862

[155] F. Valdivieso, J. Vázquez, M. A. García et al (2003), "Transcriptional activation by AP-2α is modulated by the oncogene DEK," *Nucleic Acids Research,* vol. 31, no. 5, pp. 1571-1575

[156] S. A. Myers, J. Wright, F. Zhang et al (2017), "CRISPR/Cas9-APEX-mediated proximity labeling enables discovery of proteins associated with a predefined genomic locus in living cells," *bioRxiv,* p. 159517

[157] C. Burkhart, R. Kohrn, B. Walker et al (2014), "Abstract 800: Synergistic effects of CBL0137 and gemcitabine against non-small cell lung and pancreatic cancer xenografts," *Cancer Research,* vol. 74, no. 19 Supplement, p. 800

[158] C. Burkhart, D. Fleyshman, R. Kohrn et al (2014), "Curaxin CBL0137 eradicates drug resistant cancer stem cells and potentiates efficacy of gemcitabine in preclinical models of pancreatic cancer," *Oncotarget,* vol. 5, no. 22, pp. 11038-11053

[159] A. Safina, A. A. Purmal, A. V. Gudkov et al (2016), "Anticancer drug candidate CBL0137, which inhibits histone chaperone FACT, is efficacious in preclinical orthotopic models of temozolomide-responsive and -resistant glioblastoma," *Neuro-Oncology,* vol. 19, no. 2, pp. 186-196

[160] S. Portwood, E. S. Wang, A. Safina et al (2017), "Inhibition of the Histone Chaperone, FACT (Facilitates Chromatin Transcription) By the Novel Small Molecule CBL0137 (Curaxin) Shows Promising Efficacy in Pre-Clinical Models of Acute Myeloid Leukemia," *Blood,* vol. 130, no. Suppl 1, p. 1372

[161] S. De, D. J. Lindner, C. J. Coleman et al (2018), "The FACT inhibitor CBL0137 Synergizes with Cisplatin in Small-Cell Lung Cancer by

Increasing NOTCH1 Expression and Targeting Tumor-Initiating Cells," *Cancer Research,* vol. 78, no. 9, p. 2396.

[162] W. Wu, K. He, Q. Guo et al (2019), "SSRP1 promotes colorectal cancer progression and is negatively regulated by miR-28-5p," *Journal of Cellular and Molecular Medicine,vol.* 108, no.2, pp. 734-738

[163] ClinicalTrials.gov, "Study of IV CBL0137 in Previously Treated Hematological Subjects," https://clinicaltrials.gov/ct2/show/ NCT02931110. [Accessed on 02.03.2018]

[164] A. V. Gasparian, C. A. Burkhart, A. A. Purmal et al (2011), "Curaxins: anticancer compounds that simultaneously suppress NF-κB and activate p53 by targeting FACT," *Science translational medicine,* vol. 3, no. 95, pp. 95ra74-95ra74

[165] K. S. Bocchinfuso Wayne P. and Korach (1997), "Mammary Gland Development and Tumorigenesis in Estrogen Receptor Knockout Mice," *Journal of Mammary Gland Biology and Neoplasia,* vol. 2, no. 4, pp. 323-334

[166] Milioli H. H., Vimieiro R, et al (2016), "Iteratively refining breast cancer intrinsic subtypes in the METABRIC dataset," *BioData Mining,* vol. 9, no. 1, p 2

APPENDIX A

List of proteins after the elimination of common proteins between IgG and the triplicates of Site A samples			
UniProt	**Gene ID**	**UniProt**	**Gene ID**
P11388	TOP2A	**P06733**	ENO1
O00567	NOP56	**P61313**	RPL15
Q9UKM9	RALY	**P46781**	RPS9
Q8IY81	FTSJ3	**P67936**	TPM4
O60832	DKC1	**P27348**	YWHAQ
O43290	SART1	**P50914**	RPL14
Q9BY78	RNF26	**P46776**	RPL27A
Q9Y3Y3	SH2B1	**P0DMV9**	HSPA1B
P46013	MKI67	**P01859**	IGHG2
Q9Y2X4	SH2D3A	**Q14137**	BOP1
P82980	RBP5	**P46087**	NOP2
Q08945	SSRP1	**O75531**	BANF1
Q9Y5S1	TRPV2	**Q8TDN6**	BRIX1
Q05519	SRSF11	**P98179**	RBM3
Q9Y5B9	RBM8A	**P09211**	GSTP1
Q13186	SPTAN1	**P12270**	TPR
P27694	RPA1	**P61513**	RPL37A
P26583	HMGB2	**Q9Y2L1**	DIS3
P35659	DEK	**P62306**	SNRPF
P09874	PARP1	**Q2TAY7**	SMU1
Q92945	KHSRP	**Q08J23**	NSUN2
P52597	HNRNPF	**O00541**	PES1
P11387	TOP1	**Q9BY42**	RTFDC1
Q99729	HNRNPAB	**Q13148**	TARDBP
O43707	ACTN4	**P13984**	GTF2F2
P62750	RPL23A	**O75475**	PSIP1
O43143	DHX15	**O95232**	LUC7L3
P29692	EEF1D	**P68371**	TUBB4B
Q9UQ35	SRRM2	**P07900**	HSP90AA1
P18621	RPL17	**Q7Z794**	KRT77

Appendix table A1: List of the proteins after the elimination of common proteins between IgG and the triplicates at Site A.

Table shows the proteins remained after the elimination of common proteins between IgG control and the triplicates of Site A samples, presented in UniProt and Gene ID.

List of proteins after the elimination of common proteins between IgG and triplicates at Site B					
UniProt	**Gene ID**	**UniProt**	**Gene ID**	**UniProt**	**Gene ID**
Q09666	AHNAK	**P35637**	FUS	**P83731**	RPL24
Q86V81	ALYREF	**O75367**	H2AFY	**P61353**	RPL27
O00148	DDX39A	**P19105**	MYL12A	**P55884**	EIF3B
P68032	ACTC1	**Q02218**	OGDH	**P05388**	RPLP0
Q9NZM1	MVOF	**A0M8Q6**	IGLC7	**P62750**	RPL23A
P12956	XRCC6	**P40227**	CCT6A	**O95486**	SEC24A
P09651	HNRNPA1	**P04083**	ANXA1	**P21399**	ACO1
P55795	HNRNPH2	**P49368**	CCT3	**Q9NQ29**	LUC7L2
P23396	RPS3	**P38159**	RBMX	**P56537**	EIF6
P17844	DDX5	**P04075**	ALDOA	**P78371**	CCT2
P11021	HSPA5	**P06753**	TPM3	**Q13242**	SRSF9
Q03426	MVK	**P62241**	RPS8	**P30050**	RPL12
P34932	HSPA4	**P17987**	TCP1	**O15371**	EIF3D
O94979	SEC31A	**O00299**	CLIC1	**Q01130**	SRSF2
P13647	KRT5	**P55209**	NAP1L1	**Q15785**	TOMM34
Q14103	HNRNPD	**Q99873**	PRMT1	**P04792**	HSPB1
P62333	PSMC6	**Q86V48**	LUZP1	**P21589**	NT5E
Q96S55	WRNIP1	**P55072**	VCP	**Q13547**	HDAC1
P62937	PPIA	**P14625**	HSP90B1	**P62249**	RPS16
Q9UHX1	PUF60	**P15880**	RPS2	**P61326**	MAGOH
P42167	TMPO	**Q9H0S4**	DDX47	**P27797**	CALR
P18124	RPL7	**P36957**	DLST	**P48643**	CCT5
O60506	SYNCRIP	**O75643**	SNRNP200	**O75400**	PRPF40A
P63241	EIF5A	**Q15393**	SF3B3	**Q15365**	PCBP1
P62424	RPL7A	**P42704**	LRPPRC	**P07195**	LDHB
P60660	MYL6	**P49454**	CENPF	**P62081**	RPS7
Q8ND24	RNF214	**P53396**	ACLY	**P27635**	RPL10
O43166	SIPA1L1	**P17980**	PSMC3	**P26358**	DNMT1
Q9Y230	RYVBL2	**P62906**	RPL10A	**P23526**	AHCY
P13639	EEF2	**P38646**	HSPA9	**P04844**	RPN2
Q96J42	TXNDC15	**P12268**	IMPDH2	**O14980**	XPO1
P50990	CCT8	**Q9Y265**	RUVBL1	**P63173**	RPL38
P00558	PGK1	**P24534**	EEF1B2	**P51114**	FXR1
P63104	YWHAZ	**Q2TAZ0**	ATG2A	**Q53F19**	NCBP3
Q99848	EBNA1BP2	**P36578**	RPL4	**P28070**	PSMB4
P52292	KPNA2	**Q13200**	PSMD2	**O75821**	EIF3G
P48444	ARCN1	**P21291**	CSRP1	**P43686**	PSMC4

UniProt	Gene ID	UniProt	Gene ID	UniProt	Gene ID
P46778	RPL21	**P28799**	GRN	**Q96AE4**	FUBP1
Q01518	CAP1	**Q9UK45**	LSM7	**P33992**	MCM5
P31949	S100A11	**Q9Y3I0**	RTCB	**Q9NWB6**	ARGLU1
Q96FW1	OTUB1	**P25398**	RPS12	**Q6P1J9**	CDC73
P62266	RPS23	**Q969E4**	TCEAL3	**P25685**	DNAJB1
P30085	CMPK1	**O76021**	RSL1D1	**Q96I24**	FUBP3
Q16555	DPYSL2	**Q8NBT2**	SPC24	**P40121**	CAPG
Q8WXA9	SREK1	**O14818**	PSMA7	**P20290**	BTF3
P06576	ATP5F1B	**Q9UGR2**	ZC3H7B	**P25787**	PSMA2
P60900	PSMA6	**P15531**	NME1	**Q9NTK5**	OLA1
Q7L014	DDX46	**Q96QD9**	FYTTD1	**Q9Y266**	NUDC
P49756	RBM25	**Q9HB71**	CACYBP	**Q8WXF0**	SRSF12
Q8WXF1	PSPC1	**Q9BZZ5**	API5	**P46060**	RANGAP1
P31689	DNAJA1	**Q9BVP2**	GNL3	**Q9UNX4**	WDR3
P28066	PSMA5	**Q96PK6**	RBM14	**P62318**	SNRPD3
O94776	MTA2	**O43818**	RRP9	**P14678**	SNPRB
P05455	SSB	**O43172**	PRPF4	**O43670**	ZNF207
Q9H307	PNN	**P07305**	H1F0	**P46783**	RPS10
P55769	SNU13	**P43487**	RANBP1	**Q09028**	RBBP4
Q9Y3U8	RPL36	**Q99459**	CDC5L	**P39656**	DDOST
P62269	RPS18	**Q9H0A0**	NAT10	**Q8N684**	CPSF7
Q13347	EIF3I	**Q6PJT7**	ZC3H14	**Q99829**	CPNE1
P00846	MT-ATP6	**Q15427**	SF3B4	**Q9Y6A5**	TACC3
P18077	RPL35A	**P43490**	NAMPT	**Q8IYB3**	SRRM1
Q9NVC6	MED17	**Q15942**	ZYX	**Q96AG4**	LRRC59
O76003	GLRX3	**P35268**	RPL22	**P61026**	RAB10
Q14683	SMC1A	**P12081**	HARS	**P10599**	TXN
P42677	RPS27				

Appendix table A2: List of the proteins after the elimination of common proteins between IgG and triplicates of Site B.

Table shows proteins remained for further analysis after the elimination of common proteins between IgG and the triplicates of Site B samples, presented in UniProt and Gene ID.

List of proteins after the elimination of common proteins between IgG and the triplicates of Sites A and B					
UniProt	**Gene ID**	**UniProt**	**Gene ID**	**UniProt**	**Gene ID**
P21333	FLNA	**Q07666**	KHDRBS1	**Q03252**	LMNB2
O75369	FLNB	**O75083**	WDR1	**Q9BTT0**	ANP32E
Q14204	DYNC1H1	**P22087**	FBL	**P35244**	RPA3
Q13838	DDX39B	**Q9Y5B9**	SUPT16H	**P62316**	SNRPD2
P38919	EIF4A3	**Q07955**	SRSF1	**Q9NQW6**	ANLN
O75533	SF3B1	**Q08945**	SSRP1	**P31942**	HNRNPH3
P13010	XRCC5	**P13489**	RNH1	**Q86U42**	PABPN1
O15027	SEC16A	**P0C0L4**	C4A	**Q9UKV3**	ACIN1
P11940	PABPC1	**P35659**	DEK	**P62158**	CALM
Q9BY77	POLDIP3	**Q13151**	HNRNPA0	**P0DN76**	U2AF1L5
Q9Y3Y2	CHTOP	**P26583**	HMGB2	**P62273**	RPS29
P08865	RPSA	**Q13247**	SRSF6	**P61254**	RPL26
P46013	MKI67	**P63220**	RPS21	**Q14498**	RBM39
Q13263	TRIM28	**Q16629**	SRSF7	**Q9Y5S9**	RBM8A
Q15233	NONO	**Q9Y3B4**	SF3B6	**Q05519**	SRSF11
P06899	HIST1H2BJ	**Q01105**	SET	**Q15717**	ELAVL1
Q92841	DDX17	**Q9Y2W1**	THRAP3	**Q14847**	LASP1
P82979	SARNP	**P62847**	RPS24	**Q13185**	CBX3
Q12906	ILF3	**P84103**	SRSF3	**P0C0S5**	H2AFZ
P09429	HMGB1	**Q9BZE4**	GTPBP4	**Q9Y383**	LUC7L2
Q9Y2X3	NOP58	**P27694**	RPA1	**O14979**	HNRNPDL
P43243	MATR3	**Q15287**	RNPS1	**P53999**	SUB1
P67809	YBX1	**P62314**	SNRPD1	**Q12905**	ILF2
P11388	TOP2A	**P17096**	HMGA1	**P20700**	LMNB1

Appendix table A3: List of the proteins remained after the elimination of common proteins between the IgG control and the triplicates of Sites A and B.

Table shows the proteins remained after the elimination of common proteins between IgG control and the triplicates of Sites A and B, presented in UniProt and Gene ID.

Proteins with nuclear expressions at Site A	
UniProt	**Gene ID**
O00567	NOP56
Q9UKM9	RALY
Q8IY81	FTSJ3
O60832	DKC1
O43290	SART1
Q9BY78	RNF26
Q9Y3Y3	SH2B1
P46013	MKI67
Q9Y2X4	SH2D3A
P82980	RBP5
Q08945	SSRP1
Q9Y5S9	RBM8A
Q05519	SRSF11
Q9Y5B9	RBM8A
Q14137	BOP1
P46087	NOP2
O75531	BANF1
Q8TDN6	BRIX1
P98179	RBM3
P09211	GSTP1
P12270	TPR
P61513	RPL37A
Q9Y2L1	DIS3
P62306	SNRPF
Q2TAY7	SMU1
Q08J23	NSUN2
O00541	PES1
Q9BY42	RTFDC1
Q13148	TARDBP
P13984	GTF2F2
O75475	PSIP1

Appendix table A4: Proteins pulled-down from Site A with nuclear expressions.

Table shows list of proteins with nuclear expressions pulled-down in RIME from Site A, presented in UniProt and Gene ID.

Proteins with nuclear expressions at Site B					
UniProt	**Gene ID**	**UniProt**	**Gene ID**	**UniProt**	**Gene ID**
Q09666	AHNAK	**Q9H307**	PNN	**P46060**	RANGAP1
Q9NZM1	MYOF	**P55769**	SNU13	**Q9UNX4**	WDR3
P62333	PSMC6	**Q9Y3U8**	RPL36	**Q8N684**	CPSF7
Q96S55	WRNIP1	**Q9NVC6**	MED17	**Q99829**	CPNE1
Q8ND24	RNF214	**O43670**	ZNF207	**Q96AE4**	FUBP1
O75367	H2AFY	**Q9UK45**	LSM7	**P33992**	MCM5
P04083	ANXA1	**O76021**	RSL1D1	**Q9NWB6**	ARGLU1
O00299	CLIC1	**Q8NBT2**	SPC24	**Q6P1J9**	CDC73
Q86V48	LUZP1	**O14818**	PSMA7	**P25685**	DNAJB1
Q9H0S4	DDX47	**Q9UGR2**	ZC3H7B	**Q96I24**	FUBP3
P49454	CENPF	**P15531**	NME1	**Q96AG4**	LRRC59
P12268	IMPDH2	**Q96QD9**	FYTTD1	**Q8IYB3**	SRRM1
Q2TAZ0	ATG2A	**Q9HB71**	CACYBP	**P40121**	CAPG
P56537	EIF6	**Q9BZZ5**	API5	**P20290**	BTF3
Q13242	SRSF9	**Q9BVP2**	GNL3	**P25787**	PSMA2
P04792	HSPB1	**Q96PK6**	RBM14	**Q9NTK5**	OLA1
P61326	MAGOH	**O43818**	RRP9	**Q9Y266**	NUDC
O75400	PRPF40A	**O43172**	PRPF4	**Q8WXF0**	SRSF12
P26358	DNMT1	**P07305**	H1F0	**P31949**	S100A11
O14980	XOP1	**P43487**	RANBP1	**P30085**	CMPK1
Q53F19	NCBP3	**Q99459**	CDC5L	**Q8WXA9**	SREK1
P28070	PSMB4	**Q9H0A0**	NAT10	**P60900**	PSMA6
Q99848	EBNA1BP2	**Q6PJT7**	ZC3H14	**P49756**	RBM25
P21291	CSRP1	**Q15427**	SF3B4	**Q8WXF1**	PSPC1
P43686	PSMC4	**P43490**	NAMPT	**P28066**	PSMA5
Q14683	SMC1A	**Q15942**	ZYX	**O94776**	MTA2

Appendix table A5: Proteins with nuclear expressions pulled-down in RIME at Site B.

Table shows proteins with nuclear expressions from RIME analysis at Site B, presented in UniProt and Gene ID.

Common proteins with nuclear expressions from Sites A and B	
UniProt	**Gene ID**
Q9BY77	POLDIP3
Q9Y3Y2	CHTOP
P46013	MKI67
P82979	SARNP
Q9Y2X3	NOP58
P11388	TOP2A
Q9Y5S9	RBM8A
Q05519	SRSF11
Q13185	CBX3
Q9Y5B9	SUPT16H
Q08945	SSRP1
P35659	DEK
P26583	HMGB2
Q9BZE4	GTPBP4
P27694	RPA1
P17096	HMGA1
Q9NQW6	ANLN
P31942	HNRNPH3
Q86U42	PABPN1
Q9UKV3	ACIN1
P62273	RPS29
Q03252	LMNB2
Q9BTT0	ANP32E
P35244	RPA3

Appendix table A6: Common proteins between Sites A and B with nuclear expressions.

Table shows common proteins pulled-down from Sites A and B in RIME with nuclear expressions, presented in UniProt and Gene ID.

Proteins from Crapome database (>100 studies) at Site A			
UniProt	**Gene ID**	**UniProt**	**Gene ID**
O00567	NOP56	**P09211**	GSTP1
Q9UKM9	RALY	**P12270**	TPR
Q8IY81	FTSJ3	**P61513**	RPL37A
O60832	DKC1	**Q9Y2L1**	DIS3
O43290	SART1	**P62306**	SNRPF
Q9BY78	Rnf26	**Q2TAY7**	SMU1
Q9Y3Y3	SH2B1	**Q08J23**	NSUN2
P46013	MKI67	**O00541**	PES1
Q9Y2X4	SH2D3A	**Q9BY42**	RTFDC1
O75475	PSIP1	**Q13148**	TARDBP
P82980	RBP5	**P13984**	GTF2F2
Q08946	EZR	**P27694**	RPA1
Q9Y5S1	TRPV2	**P26583**	HMGB2
Q05519	SRSF11	**Q14137**	BOP1
Q9Y5B1	PPA1	**P46087**	NOP2
Q13186	SPTAN1	**O75531**	BANF1
P11388	TOP2A	**Q8TDN6**	BRIX1
P98179	RBM3		

Appendix table A7: Proteins with less than 100 studies from the Crapome database at Site A.

Common proteins from the triplicates of Site A in RIME analysis with less than 100 studies from the Crapome database, presented in Uniprot and Gene ID.

Proteins from Crapome database (>100 studies) at Site B					
UniProt	**Gene ID**	**UniProt**	**Gene ID**	**UniProt**	**Gene ID**
Q09666	AHNAK	**Q9NVC6**	MED17	**O43172**	PRPF4
Q9NZM1	MYOF	**O43670**	ZNF207	**P07305**	H1F0
P62333	PSMC6	**Q9UK45**	LSM7	**P43487**	RANBP1
Q96S55	WRNIP1	**O94979**	SEC31A	**Q99459**	CDC5L
Q8ND24	RNF214	**O43166**	SIPA1L1	**Q9H0A0**	NAT10
O75367	H2AFY	**P00558**	PGK1	**Q6PJT7**	ZC3H14
P04083	ANXA1	**Q02218**	PGDH	**Q15427**	SF3B4
O00299	CLIC1	**P36957**	DLST	**P43490**	NAMPT
Q86V48	LUZP1	**O95486**	SEC24A	**Q15942**	ZYX
Q9H0S4	DDX47	**P21399**	ACO1	**Q96AG4**	LRRC59
P49454	CENPF	**O15371**	EIF3D	**Q8IYB3**	SRRM1
P12268	IMPDH2	**Q15785**	TOMM34	**P40121**	CAPG
Q2TAZ0	ATG2A	**P21589**	NT5E	**P20290**	BTF3
P56537	EIF6	**P23526**	AHCY	**P25787**	PSMA2
Q13242	SRSF9	**P04844**	RPN2	**Q9NTK5**	OLA1
P04792	HSPB1	**P51114**	FXR1	**Q9Y266**	NUDC
P61326	MAGOH	**P48444**	ARCN1	**Q8WXF0**	SRSF12
O75400	PRPF40A	**P12081**	HARS	**P46060**	RANGAP1
P26358	DNMT1	**P61026**	RAB10	**Q9UNX4**	WDR3
O14980	XPO1	**Q01518**	CAP1	**Q8N684**	CPSF7
Q53F19	NCBP3	**Q96FW1**	OTUB1	**Q99829**	CPNE1
P28070	PSMB4	**Q16555**	DPYSL2	**Q96AE4**	FUBP1
Q99848	EBNA1P2	**P00846**	MT-ATP6	**P33992**	MCM5
P21291	CSRP1	**O76003**	GLRX3	**Q9NWB6**	ARGLU1
P43686	PSMC4	**Q9Y6A5**	TACC3	**Q6P1J9**	CDC73
Q14683	SMC1A	**P39656**	DDOST	**P25685**	DNAJB1
P31949	S100A11	**P28799**	GRN	**Q96I24**	FUBP3
P30085	CMPK1	**O76021**	RSL1D1	**Q9H307**	PNN
Q8WXA9	SREK1	**Q8NBT2**	SPC24	**P55769**	SNU13
P60900	PSMA6	**O14818**	PSMA7	**Q9Y3U8**	RPL36
P49756	RBM25	**Q9UGR2**	ZC3H7B	**Q9BZZ5**	API5
Q8WXF1	PSPC1	**P15531**	NME1	**Q9BVP2**	GNL3
P28066	PSMA5	**Q96QD9**	FYTTD1	**Q96PK6**	RBM14
O94776	MTA2	**Q9HB71**	CACYBP	**O43818**	RRP9

Appendix table A8: Proteins after the analysis with Crapome study at Site B. Table showing common proteins from the triplicates at Site B in RIME with less than 100 studies from the Crapome database, presented in Uniprot and Gene ID.

Common proteins from Crapome database (>100 studies) between Sites A and B	
UniProt	Gene ID
Q9BY77	POLDIP3
Q9Y3Y2	CHTOP
P46013	MKI67
P82979	SARNP
Q9Y2X3	NOP58
P11388	TOP2A
Q9Y5S9	RBM8A
Q05519	SRSF11
Q13185	CBX3
Q9Y5B9	SUPT16H
Q08945	SSRP1
P35659	DEK
P26583	HMGB2
Q9BZE4	GTPBP4
P27694	RPA1
P17096	HMGA1
Q9NQW6	ANLN
P31942	HNRNPH3
Q86U42	PABPN1
Q9UKV3	ACIN1
P62273	RPS29
O15027	SEC16A
Q14847	LASP1
O75083	WDR1
P13489	RNH1
P0C0L4	C4A
Q03252	LMNB2
Q9BTT0	ANP32E
P35244	RPA3

Appendix table A9: Common proteins from the Crapome database with less than 100 studies between Sites A and B.

Table shows common proteins between Sites A and B with less than 100 studies from the Crapome database, presented in Uniprot and Gene ID.